

***Volume 12, autumn 1997***

0435697030



Statistics Netherlands

# Statistics Netherlands

## **Voorburg:**

Prinses Beatrixlaan 428  
P.O. Box 4000  
2270 JM Voorburg (Netherlands)

Telephone: . . 31 (70) 337 38 00  
Fax: . . 31 (70) 387 74 29  
E-mail: [infoserve@cbs.nl](mailto:infoserve@cbs.nl)  
Internet: <http://www.cbs.nl>

## **Heerlen:**

Kloosterweg 1  
P.O. Box 4481  
6401 CZ Heerlen (Netherlands)

Telephone: . . 31 (45) 570 60 00  
Fax: . . (45) 572 74 40

***Netherlands***  
***Official***  
***Statistics***

***Special issue***  
***EDI: The State of the Dutch Art***

***Volume 12, autumn 1997***



Statistics Netherlands, Voorburg/Heerlen, 1997

Key figure A-125/1997

© Statistics Netherlands, Voorburg/Heerlen 1997.

Quotation of source is compulsory.  
Reproduction is permitted for own use or internal use.

Obtainable from Statistics Netherlands, Voorburg.

Annual subscription: Dfl 40.00  
Price per copy: Dfl 20.00

ISSN 0920-2048

Postage will be charged.

## ***Contents***

---

<b>Foreword</b>	5
EDI in the collection of statistical data: an introduction <i>Winfried F.H. Ypma, Ad J. Willeboordse and Wouter J. Keller</i>	7
Cryptography for secure EDI <i>Jan W.P.F. Kardaun and Leon C.R.J. Willenborg</i>	16
Electronic data communication modes in statistical practice <i>Maarten Boon and Dennis Ramondt</i>	24
An overview of EDI projects at Statistics Netherlands <i>Michel M. Beekman and Hens M.T. Lunter</i>	28
CBS-IRIS: an EDC tool for international trade statistics <i>Vick H.M. Smeets</i>	41
The TELER-EDISENT project <i>Gerrit W. de Bolster and Kees J. Metz.</i>	51
Electronic supply of data for labour statistics <i>Freek Arnoldus</i>	60
EFLO: Electronic exchange of local government financial data <i>Fred G.J. Arkesteijn</i>	69
VIS as secondary EDI source <i>Geert Bruinooge, Frans P.M.M. Nijsten, Hen J.M.V. Pustjens and Eric Smeets</i>	78
Electronic commerce and EDI: a European perspective <i>Philippe Lebaube and Maarten Boon</i>	86
EDI, the future <i>Wouter J. Keller</i>	100

*Editor in chief:*  
Jeroen Pannekoek

*Guest editor:*  
Winfried F.H. Ypma

*Coordinating editor:*  
Lieneke E. Hoeksma

## ***Foreword***

---

As in many other areas, in the field of statistics there is an increasing demand for electronic communication facilities. Not only from users, who can retrieve and combine data more easily with computers, but also from the respondents who supply the basic information for these statistics, as it is easier for them to transfer data by diskette or e-mail than to painstakingly complete official questionnaires by hand. For this reason, but also because technological progress has its own momentum, which cannot not easily be disregarded, Statistics Netherlands is currently making an effort to introduce *Electronic Data Interchange* or *EDI*.

This special issue of Netherlands Official Statistics is devoted to the state of affairs of EDI as used for the collection of statistical data by Statistics Netherlands. The articles in this issue focus mainly on statistical aspects rather than on technique.

The first, introductory, article presents the relevant terminology and a possible way to classify EDI projects. This is followed by two more or less technical articles – one on security and confidentiality, the other dealing with technical modes of communication – and a comprehensive overview of current EDI projects at Statistics Netherlands.

The greater part of this special issue discusses five specific EDI projects. The first is IRIS on intra-EU trade, one of the first EDI projects at Statistics Netherlands to address a large group of respondents. As it is based on – user friendly – data entry into an electronic questionnaire, its aims are still modest.

TELER/EDISENT is a project primarily directed at companies' financial administration. With support from Statistics Netherlands, respondents themselves have to define the technical and conceptual translation of their data to meet the statistical requirements of Statistics Netherlands. Because of the diversity of the administrations each respondent has a unique translation scheme.

EGUSES is designed to take data from company payroll administrations. Although the diversity of the administrations is less in this area, and the information required is closer to the payroll concepts, respondents still have to define the translation scheme, helped of course by Statistics Netherlands where necessary.

The EFLO project collects data from local government. Here only a technical translation is carried out into a syntax understandable by Statistics Netherlands. This translation is incorporated in the software package the respondent uses. Once installed the effort by the respondent is very small. The conceptual translation is done at Statistics Netherlands.

Lastly, the VIS project, which addresses neither enterprise nor institution individually, but instead uses information from another source: the Internal Revenue Service. Obviously any necessary translation is done by Statistics Netherlands. Here the response burden for the units described is eliminated altogether.

The last two articles are more general. First an overview of EDI from a European perspective. The final article rounds it all off, discussing not only future developments of EDI itself, but also the consequences EDI has for the statistical production process.

Winfried F.H. Ypma

# EDI in the collection of statistical data: an introduction

Winfried F.H. Ypma, Ad J. Willeboordse and Wouter J. Keller

*This introduction looks at the concept of EDI in the context of the collection of statistical data. One of the differences between EDI and 'paper' data collection is that EDI, by its technical nature, is much more formal. It provides us with an opportunity to analyze the process and to understand the typical problems of this type of data collection. The question 'who translates the concepts from one system (business) to the other (statistics)?' will prove to be very important. A possible classification of EDI is presented in which this question plays a role. A major reason for the introduction of EDI was the growing opposition to the so-called response burden. Therefore the relation between the different types of EDI and the administrative burden is shown. This article leaves the technical problems aside.*

The term Electronic Data Interchange (EDI) is commonly used by technicians to indicate any transfer of data between two automated systems through an electronic link, mostly by (telephone) cable. In the context of the collection of statistical data, and indeed of their dissemination, EDI is used to indicate the phenomenon that data are electronically transferred from one automated system (the respondents' computer) to another (the statisticians' computer), whatever the technical means of communication. The essence is that there is no time consuming data entry on a paper questionnaire. The very elimination of the paper questionnaire was one of the driving forces behind the introduction of EDI in statistical data collection.

For most cases, the word *interchange* is something of a misnomer, as the data traffic is usually one way in the process of data collection. The term *Electronic Data Capture* would be more appropriate. The present issue of Netherlands Official Statistics is not very strict on this point and uses both abbreviations: EDI and EDC.

## 1. Pull and push towards EDI

### *Demand pull*

There is currently a strong political call for a reduction of the response burden as part of alleviating the overall administrative burden of enterprises. Until recently Statistics Netherlands sent out 1.25 million questionnaires a year to businesses and other institutions. Large and medium-sized enterprises may receive as many as 50



questionnaires a year, including repetitive monthly and quarterly surveys. Large manufacturing companies in particular may be approached for as many as twenty different surveys. The conclusion is obvious: Statistics Netherlands has to *fight the form-filling burden*. Furthermore, as budgets are shrinking there is a demand for higher efficiency and higher productivity. We are confronted by demands for user-friendlier output formats. One particular aspect is a demand for an improvement in the coherence of the totality of the information we offer. Another aspect is that our clients will want to be able to use the new media IT has to offer.

#### *Technology push*

On the other hand we are blessed with information-technological (IT) developments or the technology push.

In the first place these developments give us new technical possibilities, the means to construct new tools for our production process. We see large improvements in the possibilities of data processing, data storage and data transmission. The latter aspect will probably have the most striking influence on our work: the communication of data between our respondents and the national statistical institute (NSI) on the one hand and the communication of data between the NSI and its clients on the other.

In the second place these new developments create their own demand. The new technology will be used everywhere. Our data suppliers will use it. Our clients will use it. They will no longer be satisfied with the old-fashioned – i.e. paper and mail – communication channels. Our suppliers produce their data by electronic means and will want to use these means to deliver those data directly to us, thus minimising their own costs. Our clients process our data by electronic means and they will demand to be able to select and receive these data with the tools that IT has to offer.

These two factors lead to the conclusion that in its production process, the NSI will have to make the strategic choices that make the best use of the possibilities of IT.

#### *Strategic choices*

New demands and new tools will affect all the aspects of our production process. To describe them let us first discern three stages within this production process. The input phase is where the data are collected in contact with the respondents. In the throughput phase these data are processed to produce the information with the characteristics we are actually looking for. In the output phase this information is offered to and disseminated among our clients.

Let us begin with the input side, where the data are collected. First, data collection among individuals and households. It would be no exaggeration to state that Statistics Netherlands has already taken a major step forward in this respect. We have introduced all kinds of Computer Aided Interviewing (CAI), and developed BLAISE to this end – a

program that can not only develop and present electronic questionnaires, but much, much more. The gains of these developments have mainly been in terms of an increase in productivity and efficiency. Not only was the number of staff needed for coding, data entry and checking activities dramatically reduced, but greater efficiency was also reflected in the much faster production of results. However, there is even more to gain, firstly on the efficiency of the production process itself; and secondly, in the statistical sphere, too, improvements are still possible. New ways of interviewing are being developed: CASI, computer aided self interviewing, and, though not directly a matter of IT, more efficient sample designs.

There is much more, however, to be done in the field of collecting data from enterprises. The demands here are greater. The response burden has become an issue, and indeed is the driving force behind our strategic choices here. At the same time we see that almost everywhere automation and IT have invaded the bookkeeping systems of the respondents involved, and this makes it very clear what our task for the immediate future must be: the *Edi-fication* of the collection of information from enterprises by the NSI. What CAI is for interviewing households, EDI (electronic data interchange) will be for data collection from enterprises. We shall examine EDI with enterprises in further detail later on in the present article.

In the throughput phase we are looking for more efficient ways to process our data. Of course CAI and EDI make much of the editing superfluous, and fewer errors will be made. Still we expect much from more efficient or rational ways to handle the editing process. Data processing is the key here. The option will be to no longer edit each individual record, but to use the computer to find the worst errors and to help correct them. At the same time the computer can prevent us spending time and money on correcting unimportant errors. The gains here are primarily in the area of productivity.

Lastly, the output phase, and it is here that the new developments probably receive most public attention, as new media are developed to present information to users. Although paper publications may continue to play their role, more professional users in particular will want to select and receive their data by electronic means. Statistics Netherlands is producing and developing those means, and makes data available on CD-ROM and via the Internet.

What is more important, and perhaps more difficult, is how data should be presented on these new media. As the amount of information will be much larger than we have ever published in our paper publications, the management of the meta-information will become crucial. For this purpose Statistics Netherlands is developing *StatLine*, ultimately a database intended for end users that should give access to 'all' our data. As was to be expected, structuring these data is proving to be the main problem. At the same time we are confronted by a lack of coherence due to a lack of statistical coordination. Nevertheless, we are aiming at a first, complete version of *StatLine* by the beginning of 1998.

StatLine has been assigned a key-role in the dissemination process of our data. The strategic choice has been made to aim for a structure in which all publications and all other dissemination of data goes through StatLine.

For the remainder of this article we shall focus on the input side.

## 2. The basics of EDI

The basic idea behind EDI in the collection process is quite simple and straightforward: given the fact that both respondents and the statistical office have electronic information systems, why not establish a direct electronic link between the two systems? As usual, the answer is less simple than this rhetorical question suggests. Direct communication between the two systems may be hampered by:

- technological dissimilarities, such as data structures, formats and software, as well as limited data communication;
- conceptual dissimilarities, such as:
  - \* naming and coding of data items;
  - \* level of aggregation of data items (a statistical item may be composed of different accounting items);
  - \* existence of data items (a statistical concept may have no accounting counterparts).

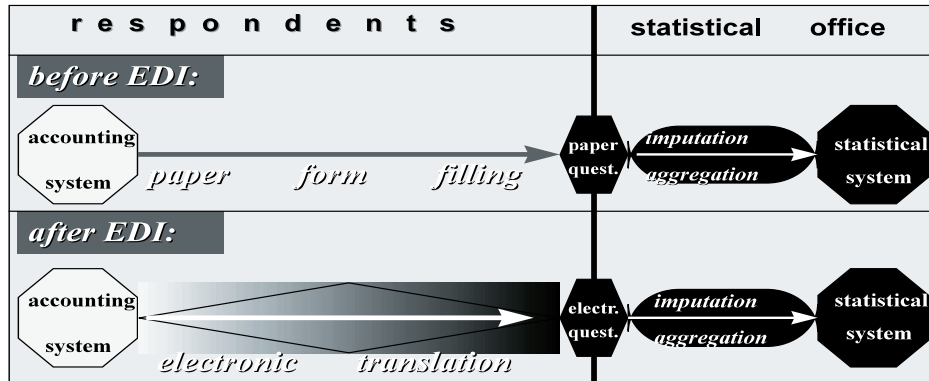
As this situation applies for most business surveys, edification of data collection is not just a question of automating paper questionnaires, but also of designing an electronic translation facility to bridge the technological and conceptual gaps between the world of the respondents and that of the statistical office.

Following the nature of the dissimilarities as mentioned above, the act of translation may refer to four different aspects:

- a *technological* aspect, i.e. transformation of data structure and format from one system to another;
- a *linguistic* aspect, i.e. the translation of names and codes from the language of the respondent to the language of the questionnaire;
- a *computational* aspect, which applies when the questionnaire data item has to be calculated from a number of accounting items and sub-items, e.g. by addition and/or subtraction;
- an *estimation* aspect, referring to questionnaire items for which there are no pure equivalents or building blocks in the respondents records (e.g. 1:n or m:n relations). The data have to be estimated on the basis of nearby or related items, to which conversion keys are applied.

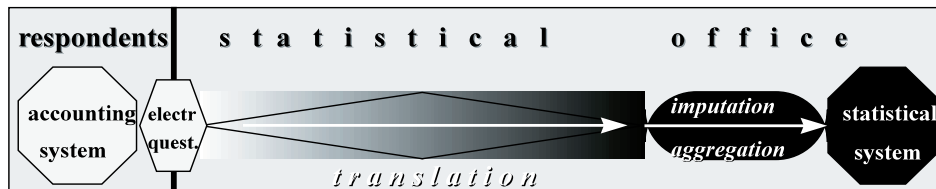
The latter three aspects actually represent the explanatory notes from the paper questionnaire, in the understanding that after edification these notes are expressed in terms of the business accounts language and in automatic translation rules. The move from paper to electronic is illustrated in Figure 1.

Figure 1



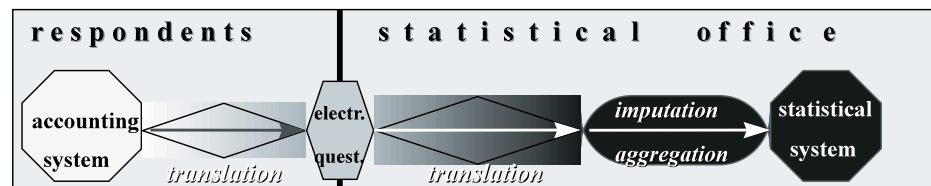
This figure depicts the situation where the statistical contents of the electronic questionnaire fully correspond to those of the paper questionnaire it replaces, under the assumption that it is indeed possible to bridge the gap electronically. In practice, however, this assumption often only holds after adaptation of questionnaire concepts in the direction of accounting concepts. If the adaptation were to amount to 100%, (electronic) translation by the respondent no longer applies. If the statistical system has to remain unchanged, this implies that the translation from accounting concepts to statistical concepts must be done by the statistical office itself (leaving aside whether or not this translation occurs electronically):

Figure 2



A typical edification project, as we shall discuss in this article, often shows a mix of both extremes:

Figure 3



Note that the statistical system has shifted slightly to the left; it is indeed not unusual for some concessions to be done towards accounting concepts. The figure illustrates in which respects the scope of edification projects tends to reach beyond information technology: the contents of the questionnaire are affected, as well as the statistical processing and the statistical system.

In summary, edification projects, while primarily oriented towards reduction of response burden, tend to entail a process of critical reconsideration with respect to the questions of whether:

- *users really need what they used to get.* Obviously, this question holds in particular for data which do not show up in business accounts. Putting forward this question is not mere form. The growing tendency to compile *reality statistics* instead of artificial constructs, has created a favourable climate for adaptation towards the perception of businesses as actors in the economic processes;
- *we should really collect all the data we need.* Calculation, imputation or estimation, either on the basis of nearby or related business accounting concepts, or obtained by confrontation with data from other surveys or by statistical integration, may be worthwhile alternatives.

In addition to this critical reflection on the *contents* of the questionnaire, a third question arises with respect to the *organisation* of data collections. The boundaries between surveys – and thus between questionnaires – are traditionally delineated without regard to how accounting systems are organised within businesses. EDI more or less forces surveying statisticians to tune their surveying strategy to the latter, as we shall see when discussing the project running at the moment. On the one hand, EDI logically demands *integration* of these different questionnaires, drawing on one and the same accounting system. This eliminates one of the most irritating response burden aspects, i.e. redundancy, caused by overlapping questionnaires. On the other hand, EDI requires *segregation* of questionnaires for which data have to be taken from different accounting and sub-accounting systems.

### 3. Basic EDI modes

The nature and complexity of edification projects will vary among surveys. The most relevant discriminating factors in this respect are:

- the distance between business accounting systems and information needs with respect to the technological and conceptual dissimilarities as mentioned above;
- the degree of standardization of business accounting practices.

When applying these criteria, three main modes of EDI emerge:

- A. *direct tapping*. This mode refers to the situation as depicted in Figure 2, where no translation is needed on the collection side. Edification consists of selecting the relevant items of the business accounts and installing communication provisions. This mode applies when the demand of the statistical office – as reflected in the questionnaire – fully corresponds with business accounts supply, which logically implies that all businesses use the same accounting system, both technically and conceptually. These ideal conditions seldom occur. A Dutch example is the survey on municipal expenditure, where the gap between statistical and accounting systems has been fully bridged by the statisticians themselves;
- B. *standardized translation*. Here one electronic questionnaire is designed, as well as one or a limited number of standard translation modules. This mode applies when there is a moderate distance between demand and supply and when accounting practices are highly standardised. Some of the Dutch labour surveys are edified according to this mode;
- C. *unique translation*. In case of non-standardized business accounts, each respondent has to establish its own unique translation module.

Although strictly speaking it does not satisfy the definition of EDI, it is nevertheless useful to add a fourth category:

- D. *semi-EDI*. Either the technological or the conceptual distance between accounting and questionnaire concepts remains too large to be bridged by automatic translation: the electronic questionnaire is filled in manually.

The sequence of this list is not coincidental: the higher the ranking, the more favourable the impact on response burden. Therefore, edification projects should in principle aim at attaining the highest possible level in the list.

The above list is based on the criterion of how the conceptual gap between business records and statistical information is bridged. Another division of EDI projects is also used at Statistics Netherlands, based on the source of the data.

We define the process as *primary* EDI if the unit to be described is also the unit that delivers the data. This is always the case when paper questionnaires are used. The

enterprise to be described also fills in the questionnaire or takes care of the electronic delivery of data. This can also be done indirectly, if for example the respondent orders another party – an accountant's office, say – to take care of responding to the business surveys of Statistics Netherlands. The original respondent remains responsible and will also pay the accountant for taking care of answering the survey.

*EDI is secondary* when the source of the information consists of data collections administrated by other – official – institutions like the tax office, the social security, organisations chambers of commerce etc. We use the information as it is available. The units to be described are no longer involved, nor are they responsible to Statistics Netherlands for the data. Clearly, secondary EDI creates no response burden for the individual enterprises. The disadvantages, however, are also evident. An arrangement has to be made with the administrator of the data, sometimes involving extra costs. The most important problem of course is that statisticians have little influence on the contents of the databases involved; they do not design the 'questionnaire' for the corporate tax.

#### **4. EDI and the response burden**

One of the main reasons for Statistics Netherlands to put so much effort into EDI was to cut back the response burden. The EDI described under A in the previous section will reduce the response burden to practically zero. For mode B, it stands to reason that the statistical office will provide respondents with standard software packages. If these are provided free of charge – which is indeed the policy of Statistics Netherlands – here too the burden will practically be eliminated. With mode C, the lack of standardisation makes it inevitable that the respondents will have to do the job themselves. Naturally, this burden is of quite a different nature than the paper burden it replaces. Recurrent efforts for questionnaire completion are replaced by one initial investment to implement the translation rules, followed by – probably minor – recurrent maintenance activities. The complexity of the rules and the stability of the accounting practices and of the statistical system will determine the size of the remaining response burden. Although it is difficult to estimate in advance, Statistics Netherlands is confident that the remaining burden will amount to a mere fraction of the paper burden.

As stated above, the response burden will also disappear if it is possible to substitute secondary EDI for 'paper' surveys. The problem is of course that not all the information can be found in the sources available for secondary EDI.

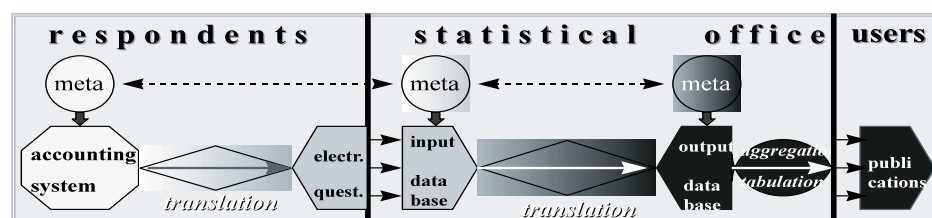
## 5. EDI and meta-data

The figures show that after edification the data flow from accounting to statistical systems will become significantly more disciplined. In order to establish reliable translation rules, the respondent must be informed precisely about the meaning of the data items he has to report on. Therefore, definitions should leave no room for deviating interpretation; questionnaire items must be defined exhaustively, i.e. in terms of *inclusions* and *exclusions* of items from the bookkeeping records. Moreover, there may be a need for a variety of questionnaire types, each designed to fit in with the language of a specific, homogeneous group of respondents.

These complex relationships on the one hand, and the growing need for *coherent* statistical data over the whole range of business statistics on the other, demand highly disciplined and coordinated data processing. A central *input meta-database*, listing all relevant concepts and their definitions, is an indispensable tool to manage and control these processes effectively.

Actually, we need *two* meta-databases: alongside the *input* meta-database to define *questionnaire* concepts in terms of accounting language, we need an *output* meta-database to define *publication* concepts in terms of questionnaire items. Where the former supports the respondent in establishing *his* translation rules, the latter supports the statistician in setting the rules for translation of input data to output data.

Figure 4



Thus, the translation is done not only by the respondent, but also by Statistics Netherlands. Each takes its share in bridging the gap. As the choice of questionnaire concepts eventually determines how the total burden is distributed between respondent and statistical office, designing the questionnaire can be seen as a trade off. Obviously, our policy is to take responsibility for the largest part and to leave the bits that can easily be converted into automatic rules to the respondent. This policy applies particularly for the estimation aspect of the translation. Here, response burden is not the only reason we prefer to do the job ourselves: the possibility of control is also a strong argument as well. However, this policy should not be taken in an absolute sense: if a specific translation activity can be done better and at less cost by the respondent himself, this is a good reason to leave the job to him.



# Cryptography for secure EDI

Jan W.P.F. Kardaun and Leon C.R.J. Willenborg

*EDI involves electronic communications. Cryptography is a convenient tool to control the authorised audience and to ensure the integrity of electronic communications. This article analyses the advantages and disadvantages of public key and secret key cryptosystems in the case of EDI, especially in the light of the consequences of key management. The analysis indicates that public key cryptosystems are preferable.*

## 1. Introduction – why cryptography?

The term *cryptography*<sup>1)</sup> conjures up somewhat dark, romantic and incomprehensible images, with stolen code books, wiretappers and burglaries all associated with tales of suspense (Kahn, 1967, Bamford, 1982). Here we hope to show that cryptography is actually nothing but a useful bag of tools to ensure that messages can only be read by the people they are intended for and that they cannot be faked. In the present era of electronic communication, including EDI, cryptography is more necessary than ever before – even if you do not have secrets to keep. As electronic communication has enabled the transfer of messages across large distances in a matter of moments, so the same technology has made it easier to automatically filter a large stream of messages to pick out the few that are of interest to an eavesdropper. In this respect EDI is more vulnerable than traditional mail. The principle of cryptography – though perhaps not the details – is easy to understand, and is in fact quite similar to what we have been doing for centuries with traditional mail: put the message in an envelope, or fold it, so that the contents are not visible, sign it and seal it.

Is cryptography really necessary for EDI between businesses and a statistical office? The answer is *yes*, for several reasons. Of course one could try to avoid the need for cryptographic protection by limiting EDI to carriers and links that can be guaranteed to be immune to eavesdropping or wiretapping. But this would not only limit the flexibility of EDI, it would probably also be far more expensive than using cryptography and assuming that the communication links are unsafe. In the case of communication via the Internet in particular, too many people can read all passing messages, and although most Internet providers are probably just as reliable as the telephone companies, the turbulent growth and development in this branch means it is easy for someone with less than honourable intentions to intrude the systems – or become a systems maintainer or even

an Internet provider. And remember: the best hackers in the world are likely to be on a government pay roll <sup>2)</sup>. The bottom line is, that even before leaked communications have done any harm – and most eavesdropping will remain well concealed – participants in an EDI system must trust the protection of the private nature of their communications enough to make participation in the EDI system appealing. As long as there is no other well accepted and widespread provision, statistical offices have to offer cryptographic protection as an integral part of its EDI systems.

## 2. Purpose

The purpose of cryptographic protection is not to provide absolute and eternal protection against unintended decryption, but to make it unattractive for non-correspondents to try to 'read other people's mail' by making the effort (in terms of time and money) required to break the encrypted message(s) greater than the contents of the messages are worth (or greater than other means of obtaining the same information). To ensure this, a cryptographic system, which consists of encryption (*E*) and decryption (*D*) algorithms, one or more keys (*K*), and a key management system (KMS), must not break down if any part of it is corrupted. To be more specific: even if the contents of one message get into the wrong hands, the other messages sent within the same cryptographic system must remain safe. The term used for this situation is that the cryptographic system must be able to withstand a plain text attack. Even if one key is revealed (e.g. the key of the day, the key of one correspondent), messages encrypted with other keys within the same system should remain safe. It is usually not realistic to rely on the secrecy of the *E* and *D* algorithms themselves – unless the number of participants is very small and each of them can be entirely trusted.

The most vulnerable points of any cryptographic system are the KMS and the proper use of the system. Therefore a cryptographic system should also be easy to use, in order to ensure user compliance with the protocols. Here we shall limit ourselves to the use of cryptography for communication, and EDI in particular. Other applications for statistical offices, such as record linkage without knowing identities, allowing clients to order and pay for publications or to browse through data bases, teleworking etc., are described elsewhere (Kardaun and Willenborg, 1995). Schneier (1994) gives a good overview of cryptography in general. Other introductory sources can be found on the World Wide Web <sup>3)</sup>. Cryptography is no longer the exclusive domain of the military and diplomatic services; it has become an economic and political issue <sup>4)</sup>. Developments are in progress that require every e-mail message, even the trivial ones, to be cryptographically protected (Zimmerman 1995, Bishop 1991).

Just to put things into perspective: the need for securing communications did not emerge only after the advent of networks, nor does the use of electronic communications justify a paranoid and over-cautious, restrictive atmosphere. We merely want to have a lock on our communication, just as we have locks on our doors.

### **3. Functions**

The basic function of a cryptosystem is protection against

- a. eavesdropping, i.e. non-correspondents reading the contents of a message;
- b. interception, i.e. the message not being delivered to its intended recipient or recipients; unnoted or unauthorised alteration of the contents of a message, especially tampering with the date and time stamps of a message, which indicate when it was written or sent;
- c. identification and authentication problems, i.e. pretending that a message has been sent by a false sender, either non-existing or existing, but not the actual sender.

Often you want all these basic functions to be activated at the same time, but not always. If you have a public message ('to whom it may concern') you will not be bothered about eavesdropping or interception, but you will require a valid signature and an unaltered text. This is why cryptography is also useful even if you have no secrets to keep ('to seal, not to conceal'). We leave it to the reader's own imagination what might happen if a communication system is not resistant to some of the possible manipulations mentioned above. In the case of EDI with statistical offices, important business information (financial and production data, investment plans, trade records, etc.) may fall into the hands of rival companies.

### **4. Communication**

For EDI applications at Statistics Netherlands the most appropriate communication model is the star topology; one central statistical office which communicates with several hundred (or thousand at a later stage) enterprises or other organisations. The communication is bi-directional. If the enterprises were to use the same cryptographic system to communicate among themselves we would have a mesh topology. This is not just a superfluous distinction, but an important characteristic, as we shall discuss below. Most of the messages are part of an EDI system proper, so they have fixed formats, and consist of communication between computer programs only. But the same secure communication channels are also be used for ad hoc queries, reminders, re-transmission of the enterprises' (own) data and general e-mail like messages, mostly between human correspondents. As each communication with the enterprises takes only a few seconds

there are no special, reserved, links, i.e. no link protection. Consequently the transmission medium has to be considered unsafe and therefore one is free to choose and change the actual transmission system.

## 5. Encryption and decryption algorithms

Two kinds of  $E$  and  $D$  algorithms are currently in use. These are not particular algorithms, of which there are many, but two families of encryption algorithms. They differ in whether the same key (and a mirrored algorithm) is used for encryption and decryption, which makes these two processes symmetrical, or whether encryption uses a different key than decryption (and consequently asymmetrical  $E$  and  $D$  algorithms are used). The two different keys have to be members of a pair, not just any two keys. As stated before, the algorithms are supposed to be non-secret, therefore the symmetrical systems rely on keeping the (single) key secret among all allowed correspondents. In the asymmetrical case, one of the keys of the pair is kept secret, while the other is intentionally made public. Therefore, symmetrical and asymmetrical cryptosystems are also called *secret key* and *public key* cryptosystems respectively. *Public key cryptosystem* is in fact something of a misnomer as this also involves a secret key. Data Encryption Standard – or DES for short – is a well-known secret key cryptosystem; and RSA (after Rivest, Shamir and Adleman, the developers of the underlying algorithm) a well-known public key system<sup>5)</sup>.

There has been a lively debate in the past about the relative merits of each of these two classes of cryptosystems, concentrating on issues like speed, patents, level of security and the role of government, for which the reader is referred to the weblinks mentioned above. Here, we shall elaborate on only one aspect of the differences between these two classes of cryptosystems: key management. For the other issues we only remark that both classes are equally safe (if properly used), can reach sufficient speed, and that a number of well-matured, specific algorithms are available.

Just a word about how asymmetrical cryptosystems work (how symmetrical systems work is rather intuitive). In order to send a message  $M$  to a recipient  $R$ , it is necessary to know  $R$ 's public key. *Anybody* can send such a message to  $R$ , precisely because the required key is public. Only  $R$ , however, can decrypt  $M$ , because to do so he needs the secret part of his particular key pair. If  $R$  wants to ascertain that the message has really been sent by the presumed sender  $S$ ,  $M$  (or a short summary of it) must be encrypted with the  $S$ 's secret key.  $R$  can verify this by decrypting with  $S$ 's public key, which he has or can obtain. Although this may seem a bit complicated and awkward, it has been designed so as to make keys belong to one party in the communication, and not to a pair (or pool) of correspondents.

## 6. Key management

One of the most important and vulnerable aspects of a cryptosystem is the key management. It is also a non-mathematical aspect, easy to grasp and difficult to implement. The keys give access to the communication, and just like physical door keys, after a while too many copies turn out to be in circulation unless strict measures are taken. If keys are frequently replaced, for example daily, you need to agree on many keys, and the synchronised replacement is prone to mistakes. If keys are kept on a computer, this computer, its backups and all computers connected to it have to be guarded. If a key is put into a chip card it is vulnerable to loss or theft. If it is only in somebody's memory he would have to be immortal. Not only should (real) keys be kept secret but they should be distinguishable from fake keys and immune to manipulation. Introducing a pseudo-key in a system under attack may give marvellous opportunities to unravel the cryptosystem used.

In spite of these pessimistic remarks, protocols have been devised offering resistance to the more common mishap and to human error. But first let's see which key management events exist in the EDI context, even if no mishap or error occurs:

- a. A key has to be issued to each correspondent (asymmetrical systems) or to a pair or group of correspondents (symmetrical systems).
- b. The key has to be sent to the other parties, either by a secure channel (symmetrical) or in other ways to reach agreement on the authentic status of the key.
- c. As many enterprises (=correspondents) have a limited life span, keys have to be revoked.
- d. During the lifetime of one enterprise, keys have to be replaced – for example daily, or weekly, or even every five years – in order to compartmentalise communications, if key-handling employees leave or go to work for a business rival, if enterprises merge, or split up, if a key gets lost (computer crash, forgotten) or disclosed.
- e. Keys that are no longer in use have to be protected against disclosure if previous communication is still of a sensitive nature.

It is in the key management that symmetrical and asymmetrical cryptosystems differ most. Assume that we have  $N$  participants, of whom  $M$  pairs or subgroups exchange secret messages. We can view this as a graph, in which there are  $N$  nodes (*communicators*) and  $M$  edges (*communication links*). We shall call this graph a secure communication network. If such a network is a star,  $M=N-1$ . If each participant corresponds with every other participant, the network is a complete graph, and the number of links  $M=N(N-1)/2$ , i.e. quadratic.

### *Symmetrical cryptosystem*

Each edge in the secret communication network needs a (secret) key. Since there are  $M$  edges, a total of  $M$  (secret) keys are needed. Initially, communication will be in a star network: the central node (representing the statistical office) needs to store  $N-1$  (secret) keys, and each peripheral node (representing an enterprise) needs to store just one secret key. If a key becomes disclosed the corresponding communication link becomes unsafe. A link may become unsafe through either of the two participants involved.

For example, suppose that there are 500 participants in the EDI project. If communication takes place in a star network only, the number of keys required would be 499. Later, if all participants communicate with each other, the number of keys is 124,750. Even if each participant only communicates with half of the other participants, the number of keys required would still be 62,500.

### *Asymmetrical cryptosystem*

Each node in the secret communication network has its own key pair consisting of a secret and a public key, so there are  $N$  key pairs in total. Each node holds its own secret key, together with the public keys of the nodes with which a secure communication link exists, varying between 1 and  $N-1$ . The total number of (secret and public) keys that have to be managed is  $2N$  (i.e. in our example only 1,000). The central node in particular (here: the statistical office) does not know the secret key of any of the enterprises. Therefore the disclosure of a secret key can only be blamed on the owner of this key.

So from this we see that an asymmetrical cryptosystem has a number of advantages over a symmetrical one. It is particularly more attractive if an initial star network evolves into a mesh-type network. By initially endowing a developing EDI system with a symmetrical cryptosystem, it can be predicted that the burden of the key management will become unbearable at some point in time, and in effect will jeopardise the proper functioning of the EDI system. As EDI gains momentum, it is a great benefit to have a cryptosystem that can handle communication with multiple partners easily and that is scalable.

In the above, we have only calculated the number of keys required *at a given moment*, for  $N$  participants and  $M$  links. As  $N$  and  $M$  evolve over time, and – as we mentioned under d. above – keys have a limited lifespan, the number of keys becomes even larger. Also it has to be envisioned that more than one key per participating organisation will be necessary, for example different keys for the financial, personnel and production departments, in effect increasing  $N$ .

Lastly, as far as key management is concerned, as EDI is rather message oriented (unlike continuous communication systems), in practice for each message (i.e. each message exchange session) a one-off key is generated and transmitted together with the message, obviously encrypted with a *master key*. So the master key does not have to be changed or exchanged often. When the master key is exchanged for the first time, an additional provision is needed that this is an authentic master key. This first time authentication can be done using traditional means: sealed envelopes, signatures, verification (by telephone), personal messenger if necessary <sup>6)</sup>. In larger organisations, several *sub-master keys* may be generated for different departments, divisions, management. In an asymmetrical cryptosystem it is likely that an authority will be established whose responsibility it is to certify the correct status of the public part of the key pair, just as chambers of commerce issue certificates for enterprises, or municipal administrations birth certificates for humans. For humans, public key certification and exchange methods have been set up in research circles <sup>7)</sup> or as grass-root efforts (Luckhardt, 1997).

## 7. Conclusions

Protection of electronic communication, as in EDI for statistical offices, is necessary both to prevent abuse of sensitive information and to inspire sufficient confidence in EDI to attract participants. If more than only a few correspondents are involved, cryptographic protection is an easier and more flexible means of providing this protection than physically securing the communication links – and it is certainly much cheaper. When communication systems are in the process of being set up, as is the case for EDI, facilities for cryptographic protection should be incorporated, designed to be flexible, allowing for growth and various communication needs. This is an especially easy choice, as such flexible systems can be obtained or made for the same price as rigid systems for limited communication patterns. The key management, i.e. issue, expiry, verification, and revocation of keys, soon becomes the most burdensome part of maintaining a cryptosystem as the number of participants grows. Asymmetrical cryptosystems, also known as *public key* systems, offer a number of advantages over more traditional symmetrical systems with regard to key management. Altogether, the additional effort of providing an EDI system for statistical with a good cryptographic protection, is small compared with the total effort of designing and implementing EDI systems.

## Notes

- 1) Cryptography is the art and science of keeping messages secure. Cryptanalysis is the art and science of breaking encrypted messages. Cryptology embodies both cryptography and cryptanalysis.
- 2) That does not necessarily mean they are the best paid hackers in the world. But then, in compensation, they are likely to have excellent equipment.
- 3) *good FAQ*: <http://www.phys.s.u-tokyo.ac.jp/local/other-faq/sci.crypt/1-e.html> (includes selected bibliography);  
*cryptographers' homepages*: <http://www.cs.ccu.edu.tw/~victor/person.html>
- 4) *citizen's organisations*: <http://www.eff.org/> ; <http://www.cdt.org/> ; <http://epic.org/privacy/newsgroups>: sci.crypt ; talk.politics.crypt
- 5) <http://www.rsa.com/> (includes a 200 p. introductory book).
- 6) We skip the point that the generation of the (secret, public) key pair must be done exclusively by each enterprise, without the help of Statistics Netherlands. Yet, it has to be very easy, error-proof, and above all, secure!
- 7) *PGP public key server*: <http://www.surfnet.nl/pgp/pks-toplevel.html>

## References

- Bamford, J. *The Puzzle Palace*. Houghton Mifflin, Boston 1982.
- Bishop, M. Privacy Enhanced Mail. In: Feigenbaum, J. and M. Merrit (eds.). *Distributed Computing and Cryptography*, p. 93–106. Am. Math. Society, 1982.
- Kahn, D. *The codebreakers: The story of secret writing*. Macmillan, New York, 1967.
- Kardaun, J.W.P.F. and L.C.R.J. Willenborg. Cryptological applications in official statistics. Seminar on New Technologies and Technologies for Statistics, Bonn, 20–22 November 1995.
- Luckhardt, N. Zeichensetzung; c't startet Krypto-Kampagne. *C't Magazin*, no. 4, p 32, 1997.
- Schneier, B. *Applied cryptography*. Wiley, New York, 1994.
- Zimmerman, P. *The Official PGP User's Guide*. MIT Press, Cambridge, 1995.



# Electronic data communication modes in statistical practice

*Maarten Boon and Dennis Ramondt*

*As explained in the introduction of this special issue, EDI is used for the automated response to statistical surveys. This article outlines the consequences of the various communication modes using examples of applications described elsewhere in this issue. The first section deals with a large scale concept: the CBS-IRIS tool for foreign trade statistics. The second section discusses a new development: EDINET as the Internet umbrella for EDISENT. In the last section we put forward the question of whether 'multi-mode' solutions should be considered as the best – or intermediate – choice for the coming years.*

## **1. CBS-IRIS: 7,000 diskettes and 2,000 electronic messages a month**

The CBS-IRIS tool was put 'launched' at the start of the Single European Market on 1st January 1993. The package aimed to facilitate the new direct reporting obligation for intra EU traders. In its original versions CBS-IRIS did not offer the function of electronic transmission by modem. The introduction of the software was to such an extent successful that the statistical office was inundated with diskettes, so many in fact that it was hardly possible to process them all. After some initial panicky reactions we adapted our procedures and processing systems and the diskette handling system now operates smoothly. On basis of our experiences we recommend the following principles:

- it is cost effective to provide respondents with formatted empty diskettes and envelopes to send in their monthly returns. This is the only way to guarantee a consistent quality (and thus readability) of the returned diskettes; furthermore the 'receiving' statistical process can easily be recognised. And last but not least, the costs of statistical reporting for companies will be reduced because some are borne by the statistical office;
- bar code labels are ideal for the identification of each individual return on diskette. These labels can be sent to the respondents separately and identify the respondent's number and – if required – the reporting period. Another advantage is that the reminder system can automatically be carried out before the diskettes are read, and unnecessary reminders are avoided. Also, if reading problems occur, the sender is already identified.

The implementation of DATACOM for CBS-IRIS<sup>1)</sup>, a direct communication mode between the companies and Statistics Netherlands, also encountered some teething troubles. It was quite difficult to get respondents to change their diskette procedures and hardware configurations. The autonomous increase in the use of modems for business purposes (especially for financial transactions with banks) favoured the migration towards electronic data transfer of statistical returns via networks. But in spite of this, Statistics Netherlands had to repeatedly notify the companies of the new procedures before they were actually adopted there. We suppose that the main reasons for this lay in internal task allocations – the people who complete statistical questionnaires are not usually those who take decisions on hardware – and the use of old configurations for this non-core activity. The lesson to be learned is that if you want statistical returns to be sent by modem, it is best to implement this with the introduction of new software. We did this for EDISENT and found that respondents rarely prefer to send back diskettes.

The choice for a Value Added Network provider as an intermediate between the respondents and Statistics Netherlands offered several advantages. Statistics Netherlands does not need to bother with receiving and transmitting protocols and extended hardware. The companies can send in their returns via local calls, as the further transmission by DATACOM is paid for by Statistics Netherlands, and they automatically receive an confirmation of receipt.

Electronic data transmission is much more preferable than sending diskettes by post. It is cheaper because there are no handling costs at either end, it is safer and the error rate has proved much lower.

## **2. EDINET**

In 1996 the Dutch government launched the National Action Plan (NAP) for the digital highway. Funds were made available for a competition for the best innovative ideas on increasing the use of Internet. Statistics Netherlands came up with the proposal *EDISENT for Internet*, a combination of the EDISENT concept with Internet. It looked promising because it reduces the effort required by companies to respond to statistical surveys (EDISENT), and it supports companies (Internet) by answering their questions and by providing new versions of the surveys and the EDISENT module. Another benefit is that Statistics Netherlands can return statistical branch information to the companies as 'results' of the surveys. The proposal was among the competition winners and meant the start of the EDINET project.

The EDINET module has the following main functions:

- transmission of data from companies, collected with the EDISENT module, to Statistics Netherlands;

- transmission of branch information from Statistics Netherlands to companies;
- answering frequently asked questions;
- distribution of new surveys;
- distribution of updates of the EDISENT module.

As the Internet is an open environment, it was necessary to encrypt the data collected with the EDISENT module. Statistics Netherlands sends the companies the encryption keys separately on diskette, thus offering an appropriate solution for the security problems inherent in the use of the Internet. The statistical branch information for companies differs per branch. The information is either directly derived from the output database or specifically defined for a particular branch by the statistical departments. Frequently asked questions (FAQs) are widely used on the Internet. The most commonly asked questions and answers can be looked up. The FAQs focus on EDISENT and EDINET; Statistics Netherlands website contains more general FAQs. The Internet is also used to distribute new surveys and EDISENT module updates, eliminating the need for mailing letters, questionnaires and diskettes, and thus substantially reducing distribution costs.

The EDINET project has been incorporated in existing processing systems. It uses the input database for the identification of the companies, including the selection of the surveys, and the output database for the derivation of the branch information. The project is designed for a large number of respondents. There is no administration of e-mail addresses, so it is easy to manage.

The project started about a year ago. The first stage was the development of a prototype, which is still being tested by fifteen companies. A second prototype was based on a different technical approach: Statistics Netherlands wanted to test several possibilities since Internet technology is changing rapidly. The third version, the production version, is scheduled for the autumn of 1997. The final solution will offer technical features which were not available five months ago. In the course of the development we have experienced problems with the stability of the Internet. The Windows 95 platforms are doing relatively well. The main difficulty lies in the lack of standards in the Windows 3.11 platforms. Each provider has their own – sometimes unstable – software to access the Internet. Since most companies work with Windows 3.11, this platform cannot be ignored.

Statistics Netherlands fieldworkers helped to install the EDINET pilot at the companies. The guinea pig companies are quite satisfied and many other companies and organisations have shown an interest in implementing this approach to collect data by a secure method.

The use of Internet technology makes it possible to combine Statistics Netherlands information requirements with those of other government institutions. In this way it will contribute to reducing the administrative burden.

### **3. Multi-mode electronic forms**

CBS-IRIS and EDISENT/EDINET are products of the Research & Development Division at Statistics Netherlands. The former is a tailor-made package for a specific statistical application, in this case foreign trade statistics, and the latter is more generic, aimed especially at the automatic completion of statistical questionnaires from bookkeeping systems. Statistics Netherlands is now investigating the possibilities of commercial Electronic Forms products (to be completed by hand). The main advantages of such applications are the following:

- they offer the possibility of rapid conversion of existing paper forms into electronic forms;
- they offer companies a choice between several communication modes, e.g. e-fax, e-mail, traditional printout (to be transmitted by traditional fax or mail) and – of course – the Internet;
- software development costs incurred by the statistical institutes will be greatly reduced; on the commercial market standard solutions for several platforms are offered for both the servers (the statistical institutes) and the clients (the respondents).

Of course, the great disadvantage is that automatic completion of questionnaires, as realised by EDISENT, is more profitable for companies. Therefore electronic forms are more of an intermediate step than the best final solution.

#### **Note**

- <sup>1)</sup> See also the article on CBS-IRIS.

# An overview of EDI projects at Statistics Netherlands

*Michel M. Beekman and Hens M.T. Lunter*

In the following an overview is presented of the most important projects using EDI at Statistics Netherlands. The information for this overview was collected by sending questionnaires to all departments in Statistics Netherlands directly involved in economic or social statistics, asking which of their statistics use or plan to use EDI for data collection and in which projects they were indirectly involved. This survey resulted in some ninety replies, several of which overlapped as different departments often work on the same project but from different angles. The list below gives an overview of the main relevant projects. No attempt has been made to rank them in any specific order.

The main distinction is between *primary* and *secondary* EDI. There are three types of *primary* EDI:

- data entry by the respondent;
- direct extraction from the administration of the respondent, in which case the translation from administrative into statistical data is done by a software module prepared by Statistics Netherlands in cooperation with the respondent;
- copying administrative data, in which case the translation is done by Statistics Netherlands.

We speak of *secondary* EDI when organisations or companies transmit to Statistics Netherlands a set of records collected by themselves for their own purposes. Statistics Netherlands then extracts the required data from these files.

For each project the questionnaire asked:

- the *data collected*;
- the *status* of the project (in preparation, being tested, in a pilot stage or actually in operation);
- the *target population* (in terms of the activity classification);
- the *frequency* of the survey concerned;
- the *medium* used to transmit data (the Internet, data communication other than via the Internet, diskette, tape or other means);
- who carries out the *translation* of the data.

Some of these characteristics are missing from the list below. This is because in some cases the distinction between the different stages of development were not sharp enough, the target population or the frequency had not yet been decided, and in a few cases even the medium for transmission was still under discussion. Also, some projects have been combined in order to keep the list manageable and not too repetitive.

## **Primary EDI**

### **EDISENT**<sup>1)</sup>

A module translating administrative data of each respondent into standardised statistical data. Data entry is also possible.

*data collected:* Mainly data from profit and loss accounts and balance sheets in agriculture, manufacturing industry, energy supply, construction, trade, transport and other business services (all yearly); data on total sales in kind, volume and value in industry and trade (monthly).

*status:* Partly operational.

*target population:* All businesses with twenty or more employees.

*frequency:* Yearly; monthly

*medium:* Data communication other than the Internet; diskette.

*translation:* Respondent.

### **Top 100**

A special EDISENT module in preparation for the largest enterprises, tailor made for each one, to obtain as many variables as possible on a monthly basis.

### **EDINET**

An EDISENT module especially developed for the use of Internet.

### **Teler**

An international variant of the EDISENT module.

### **Synergy survey**

An EDISENT module adapted in co-operation with EnergieNed.

*status:* Currently being tested.

*target population:* A maximum of 100 units active in the production and/or distribution of electricity, gas and water.

*frequency:* Yearly.

### **EnergieNed**

This model is also built on EDISENT concepts.

*data collected:* Data to be collected include data on the production (kind, volume, value), transformation, distribution and use of all kinds of energy and water, including the volume of capital goods used.

*status:* Still in a preparatory stage.

*frequency:* Yearly, quarterly or monthly depending on the type of information.

### **Agriculture**

This project is in fact a copy of the EDISENT module.

*status:* The project is in the pilot stage. It is to be mailed to 12 accountants offices. In the production stage data on 3,000 clients, distributed over these 12 offices will be collected.

### **IRIS**<sup>1)</sup>

A module used by the respondents to translate their intra-EU trade data into data for statistical use and for customs. The module is based on data entry.

*data collected:* Intra EU imports and exports of goods.

*status:* Operational.

*target population:* All businesses whose exports exceed NLG 500,000.

*frequency:* Quarterly, monthly.

*medium:* Data communication other than the Internet; diskette.

*translation:* Respondent.

### **EFLO**<sup>1)</sup>

Data capture from the financial administrations of local governments.

*data collected:* Revenues, expenditure and budgeting by function of government.

*status:* Operational.

*target population:* All municipalities, provinces, polders and corporations thereof.

*frequency:* Quarterly.

*medium:* Data communication other than the Internet; tape.

*translation:* Statistics Netherlands.

### **Eguses**<sup>1)</sup>

Modules used to transform administrative business data on employees into statistical data.

*data collected:* Main components of wages and other data by employee.  
*status:* Operational.  
*target population:* All businesses.  
*frequency:* Quarterly; monthly.  
*medium:* Diskette; tape.  
*translation:* Respondent and Statistics Netherlands

### **GBA**

Data capture from municipalities about their population.

*data collected:* Number of people, births, death, marriages, divorces, migration, income etc.  
*status:* Operational.  
*target population:* All municipalities.  
*frequency:* Yearly; continuous.  
*medium:* Data communication other than the Internet.  
*translation:* Respondent.

### **Consumption index**

Data capture directly from cash registers.

*data collected:* Turnover of food products by kind, volume and value.  
*status:* Operational.  
*target population:* Supermarkets and department stores.  
*frequency:* Monthly.  
*medium:* The Internet; diskette.  
*translation:* Statistics Netherlands.

### **EDI with supermarkets**

Data capture from cash registers of supermarkets. Data are used for the consumption index and the consumer price index (CPI) .

*status:* In preparation.  
*target population:* All (35) supermarket chains.  
*frequency:* Monthly.  
*translation:* Statistics Netherlands.



### **EAN Bar codes**

Related to the project EDI in supermarkets, this project comprises the development of a database of EAN bar codes and related product names.

*status:* In preparation.  
*target population:* All (35) supermarket chains.  
*frequency:* Continuous.  
*translation:* Statistics Netherlands.

### **EDI-NWR**

Data capture from the administration of the Nationale Woningraad (National Dwellings Council), the administrative centre for 425 housing associations. The data are used in the calculation of the monthly CPI and as a part of the data supply to government bodies.

*data collected:* Type of dwelling (number of rooms, square metres, type of construction etc.) rents and other costs.  
*status:* In preparation.  
*target population:* 425 housing associations.  
*frequency:* Yearly.  
*translation:* Statistics Netherlands.

### **Building permits issued**

Data capture from local governments. This project is divided into several sub-projects depending on type of source, data and/or statistics. Each sub-project is in its own stage of development. Communication with Statistics Netherlands depends on the respondent's facilities. One special facility is GEMNET, a kind of intranet among municipalities, some government departments and a few subscribers, of which Statistics Netherlands is one.

*data collected:* Building permits issued, construction output.  
*status:* In preparation, pilot, testing, operational.  
*target population:* All municipalities.  
*frequency:* Monthly.  
*medium:* The Internet, data communication other than the Internet, diskette.  
*translation:* Respondent and Statistics Netherlands.

**Water pollution**

Data entry on the treatment of waste water.

*data collected:* Waste water treated, type and concentration of toxicants found etc.  
*status:* Operational.  
*target population:* Polders and other institutions treating waste water.  
*frequency:* Yearly.  
*medium:* Diskette.  
*translation:* Respondent.

**Judicial actions**

Data capture from different types of law courts. This project comprises seven parts on civil, administrative and criminal law respectively. The civil law project is still in preparation; the others operational.

*data collected:* Data on the parties involved, the facts, the stage of the trial or investigation, decisions and verdicts and jurisprudence.  
*status:* In preparation, operational.  
*target population:* Courts of law.  
*frequency:* Quarterly, monthly.  
*medium:* Diskette.  
*translation:* Statistics Netherlands.

**Police and fire services**

Data entry or transmission by different types of administrations.

*data collected:* Information on crimes, profiles of victims and suspects, fires by type, assistance by type and environmental data.  
*status:* Testing (police districts), operational (municipalities).  
*target population:* 30 police districts, all municipalities.  
*frequency:* Quarterly.  
*medium:* Diskette.  
*translation:* Respondent.

**Fraud**

Data capture from municipalities and social security institutions.

*data collected:* Types of fraud and profiles of persons involved.  
*status:* Operational.  
*target population:* All municipalities and 26 social security institutions.  
*frequency:* Quarterly.  
*medium:* Diskette.  
*translation:* Respondent.

**Care for architectural monuments**

Data capture on architectural monuments that have undergone maintenance or reconstruction in the past year.

*data collected:* Type of building, owner, amount and percentage of money subsidised.  
*status:* Operational.  
*target population:* Approximately 5,000 monuments.  
*frequency:* Yearly.  
*medium:* Diskette.  
*translation:* Statistics Netherlands.

**Secondary EDI****Collaborative data acquisition concerning business registers**

Data capture and exchange on business registers. This project aims to collect register information from the chambers of commerce, tax registers and social security institutions on a monthly, and in the future probably daily basis. It will enable or facilitate statistics on economic demography.

*data collected:* Variables such as name, address, kind of activity and size class.  
*status:* In preparation.  
*target population:* 1.3 million units.  
*frequency:* Monthly, daily.  
*medium:* Tape, the Internet.

### **VAT register**

Data capture from the internal revenue service (IRS). It is being examined whether the data could replace the surveys used for the compilation of monthly statistics on manufacturing, retail trade and restaurant and hotel trade. The VAT register is currently used for the monthly statistics on the construction industry.

*data collected:* Turnover.  
*status:* (1) Pilot, (2) operational.  
*target population:* (1) 87,000 units, (2) 50,000 units.  
*frequency:* Yearly, quarterly, monthly.  
*medium:* Tape.

### **Corporate tax information system (VIS) <sup>1)</sup>**

Data capture from the internal revenue service. The system contains data on every taxable corporation. Studies are underway to find out whether this information could replace most of the direct data collection on these variables by Statistics Netherlands.

*data collected:* Information on profit and loss accounts and balance sheets.  
*status:* Testing.  
*target population:* 290,000 units.  
*frequency:* Quarterly.  
*medium:* Tape.

### **Income information system (IIS)**

Data capture from the internal revenue service, the Ministry of housing, planning and the environment and from a register on selection and admission of students in higher education and study grants. The data enable the compilation of statistics (national and regional) on individual income.

*data collected:* Taxpayer variables, e.g. age, sex, income and wage tax, interest paid and received, rent subsidies and study grants.  
*status:* Operational.  
*target population:* All income tax payers and students over 18.  
*frequency:* Yearly.  
*medium:* Diskette, tape.

### **Stock indexes**

Data capture from the Amsterdam Stock Exchange. The data are used to calculate the 'All Share Index' and the 'Total Return Index', published daily by Statistics Netherlands.

*status:* Operational.  
*target population:* Amsterdam Stock Exchange.  
*frequency:* 4 times a day.  
*medium:* Data communication other than the Internet.

### **Insurance companies and pension funds**

Data capture from the insurance supervisory board.

*data collected:* Financial and production data.  
*status:* Testing (insurance); operational (pensions).  
*target population:* 700 (insurance); 1,100 (pensions).  
*frequency:* Yearly.  
*medium:* Tape.

### **EDI in agricultural statistics**

Data capture from several organisations in the field of agriculture, e.g. the Ministry of agriculture, nature management and fisheries, the Institute for agricultural economics, the marketing boards for livestock, meat and eggs and various other marketing boards.

*data collected:* Mainly livestock, farmland (owned, bought and sold, leased, prices and use), labour force, production and prices of livestock, meat, dairy products and crops.  
*status:* Operational.  
*target population:* 120,000.  
*frequency:* Yearly, quarterly, monthly, weekly.  
*medium:* Diskette, tape.

### **EDI in environmental statistics**

Data capture from several governmental bodies, provinces and polders.

*data collected:* Water and air pollution, production of hazardous waste and motor vehicles to be dismantled.  
*status:* Operational.  
*frequency:* Yearly.  
*medium:* The Internet, diskette, tape.

### **EDI in traffic and water transport statistics**

Data capture from mainly local or government bodies, such as port authorities and waterways departments; in some cases from customs authorities.

*data collected:* Fleet information (seaborne and inland), ship cargoes and transport routes.  
*status:* In preparation, operational.  
*frequency:* Monthly, daily.  
*medium:* Data communication other than the Internet, diskette, tape.

### **Sagitta**

Data capture from customs authorities on extra-EU trade.

*data collected:* Imports, exports and transit of goods in kind, volume and value by country and by mode of transport.  
*status:* Operational.  
*target population:* All businesses which import or export goods.  
*frequency:* Monthly.  
*medium:* Tape.

### **Geographic register and geometric coordinates**

Data capture from all municipalities, postal service, national planning agency and the topographical service.

*data collected:* Digital borders of municipalities and neighbourhoods. Addresses including data like municipality code, census tracts and districts, grid square and postal code.  
*status:* Operational.  
*frequency:* Yearly.  
*medium:* Diskette, tape, cd-rom.

### **Real estate**

Data capture from all municipalities, the Ministry of housing, planning and the environment and the land registry office.

*data collected:* Value of real estate, number, type and value of mortgages, interest rates, balance sheets and profit and loss accounts of housing corporations.  
*status:* Testing, operational.  
*target population:* All municipalities.  
*frequency:* Yearly, monthly.  
*medium:* Diskette.

### **Social work**

Data capture from the institution enterprises active in this sector.

*data collected:* Variables from the profit and loss accounts and employment.  
*status:* Operational.  
*target population:* 160.  
*frequency:* Yearly.  
*medium:* Diskette.

### **Social security**

Data capture from social security institutions.

*data collected:* Variables on employees, e.g. address, kind of job, wages; data on recipients of benefits, e.g. address; payment ratios between different kinds of social security.  
*status:* Testing.  
*target population:* All employees (8 million).  
*frequency:* Yearly, monthly.  
*medium:* Diskette, tape.

### **Legal measures**

Data capture from judicial courts, child protection councils and Ministry of Justice.

*data collected:* Extraditions (land, reason, procedure); number and kind of child protection measures; data on prisoners.  
*status:* Operational, testing (prisoners).  
*frequency:* Yearly, semi-annual.  
*medium:* Diskette.

### **Unemployment**

Data capture from employment offices.

*data collected:* Profiles of every one registered at the employment offices (approx. 1 million people).  
*status:* Operational.  
*target population:* All employment offices.  
*frequency:* Monthly.  
*medium:* Tape.

### **Hospital care**

Data capture from the National Hospitals Institute.

*data collected:* Data on employment in hospitals.  
*status:* Operational.  
*target population:* 800 hospitals.  
*frequency:* Yearly.  
*medium:* Diskette.

### **Savings**

Data capture from the Central Bank.

*data collected:* Savings according to the balance sheets.  
*status:* Operational.  
*target population:* 80.  
*frequency:* Monthly.  
*medium:* Data communication other than the Internet.

### **Hotel accommodation**

Data capture from a research and statistics institute.

*data collected:* Number of guests and overnight stays.  
*status:* Operational.  
*target population:* 300.  
*frequency:* Monthly.  
*medium:* Diskette.



**Art education**

Data capture from the art education union.

*data collected:* Items from profit and loss accounts, employment, prices, participants.  
*status:* Operational.  
*target population:* 250.  
*frequency:* Bi-annual.  
*medium:* Diskette.

**Filling stations (car fuels)**

Data capture from an accountants' office.

*data collected:* Turnover.  
*status:* Operational.  
*target population:* 300.  
*frequency:* Monthly.  
*medium:* Diskette.

**Note**

<sup>1)</sup> Described in more detail elsewhere in this issue.

# CBS-IRIS: an EDC tool for international trade statistics

Vick H.M. Smeets

*With the completion of the Single Market, customs formalities have been abolished and no longer serve as a source of statistical information on international trade within the European Union (EU). The new statistical system that has replaced the customs documents, INTRASTAT, and the need to reduce the response burden prompted Statistics Netherlands to develop IRIS. Originally designed as a data-entry tool, IRIS contains a number of features that support the respondent's work, for example when the same information has to be provided more than once. It also enables intra-EU trade information to be supplied directly from company accounts by providing specifications to software builders and by granting conformity labels to those who build software in conformity with these specifications. Based on data entry by the respondent, in terms of the introductory article in this issue IRIS can be classified as semi-EDI. The translation from administration to statistics is done by the respondent.*

## 1. Intrastat and CBS-IRIS

International trade statistics in the Netherlands have undergone major changes in the past years, both in methodology and in data collection procedures. Traditionally these statistics were compiled from information registered by the customs authorities on the movements of goods across national borders. With the completion of the Single Market on January 1st 1993, customs formalities at the borders between member states of the European Union were abolished and customs information on the movements of goods between member states was no longer available. In order to continue the statistics on the intra-community trade a new statistical system was set up within the European Community: the INTRASTAT system (EU 1991-1), which collects the relevant information directly from the companies concerned.

Before the Single Market came into effect the statistical reports were integrated in the customs declaration procedures, which were often handled by forwarding agents. The introduction of INTRASTAT shifted the reporting burden to the desks of the importing and exporting companies themselves, and smaller companies in particular experienced this as a considerable increase in their administrative burden.

For Statistics Netherlands INTRASTAT meant redesigning the statistical process for intra-trade statistics, including the development of a register of all known intra-Community operators, an obligation explicitly stated in the Council Regulation on the INTRASTAT system.

The necessity of a whole new statistical process and the bureau's obligation to reduce the administrative burden for companies concerned prompted Statistics Netherlands to develop a tool to assist the companies, who were now the mandatory providers of statistical information. This tool is called CBS-IRIS.

Another consequence of the introduction of INTRASTAT was that there were now two series of trade statistics, each with its own methodology: the INTRASTAT system for statistics on trade between member states, and statistics on international trade with countries outside the European Union. The latter are still compiled from the information collected by the customs authorities, but no longer by means of paper forms. Dutch customs have recently started an automation project known as Sagitta, which has caused a shift in data handling from paper to EDC.

Although the present article concentrates on the INTRASTAT system and the introduction of CBS-IRIS in this process, it is important to be aware of the change in the data flow between customs authorities and Statistics Netherlands when looking at the consequences of EDC for the organisation.

## **2. Reporting enterprises and thresholds**

There is a strong link in the council regulation (EU, 1991-2) between INTRASTAT and the VAT system. Information from the VAT registration was used to compile the register of importing and exporting companies and consequently the units in this register were fiscal units for the VAT. This has the advantage that VAT data can be used to check the INTRASTAT data and where necessary to adjust the total trade amount. A disadvantage of using fiscal units is that in many cases they differ from the kind of activity units used in the General Business Register, making it difficult to compare production figures (which are based on kind of activity units) with trade figures, except for the totals.

Administrative offices hired by companies to handle their INTRASTAT returns constitute a special category. These agencies are known as *third party providers*: they take care of the INTRASTAT returns for companies, but are not obliged to submit INTRASTAT returns for themselves.

INTRASTAT legislation provides for simplification and assimilation thresholds (EU, 1992-1). In 1993 Dutch enterprises whose foreign trade within the EU (arrivals and dispatches separately) was below NLG 175,000 were exempted from sending in INTRASTAT returns (the assimilation threshold), while for those whose trade was worth between NLG 175,000 and NLG 400,000 a simplified INTRASTAT return would suffice (the simplification threshold). The simplification threshold was removed in 1995, whereas the assimilation threshold was raised to NLG 400,000. On 1 January 1997 the assimilation threshold was raised to NLG 500,000.

### **3. Data elements**

The data elements required for the CBS-IRIS program are the same as those required for INTRASTAT, as defined in several Commission regulations (EU 1992-2 and 1992-3). In addition to the basic trade message: *what was it, where did it come from or go to, how much of it was there and how much did it cost*, some extra elements are required. These include such items as *presumed mode of transport, port or airport of loading or unloading, statistical procedure, nature of transaction*. Also, the main value to be filled in on the INTRASTAT returns is not the *invoice* value, but the so-called *statistical* value. This is defined for arrivals and dispatches separately. For arrivals it is established *on the basis of taxable amount for tax purposes, minus however taxes due because of release for home use, and transport and insurance costs relating to that part of the journey which takes place in the member state of arrival*. For dispatches the statistical value is established *on the basis of taxable amount to be determined for tax purposes, minus, however, any taxes deductible because of dispatch; it shall on the other hand include transport and insurance costs relating to that part of the journey which takes place on the statistical territory of the member state of dispatch*. In the case of goods derived from processing, the statistical value is established as if the goods had been wholly manufactured in the member state of processing.

The Combined Nomenclature is used to describe the goods being traded. This nomenclature, which is also used for non-EU foreign trade statistics, contains over 10,500 different commodity codes. Code lists are prescribed for the extra data elements.

### **4. CBS-IRIS: an intelligent data entry tool**

CBS-IRIS was developed as a free software tool to be distributed among internationally trading companies. This software, in essence a DOS-based data entry tool, enables companies to compile their statistical reports. The programme uses the Blaise tools (Statistics Netherlands, 1994), which perform validity – for example, Combined Nomenclature codes – and consistency checks during the data entry process.

Although CBS-IRIS was originally developed as a data entry program, one of its main goals, to reduce the response burden for companies, can only be achieved if companies can download most of the data needed for the INTRASTAT returns from their own computer administrations. A main handicap here is that many of the data required are not usually included in the companies' systems. And of the data elements that *are* available, some, for example commodity codes, have to be translated from the codes used by the company to the codes prescribed by INTRASTAT.

A number of features were built into CBS-IRIS to facilitate the compilation of INTRASTAT returns:

- a module was built in to download required data from the ASCII listings that many administrative programs can produce. This way a company can import a large amount of the data and has only to add, by way of data entry, the extra data elements;
- the international trade of many companies has consistent characteristics: they always had to enter the same values for *statistical procedure*, *nature of transaction*, sometimes even for *member state of destination* and *member state of consignment*. CBS-IRIS built in the option to define a number of default value sets. In combination with the downloading of data already available, these defaults greatly facilitate the compilation of the INTRASTAT returns;
- a translation module was incorporated, in which companies can put together a list of their own product codes and the matching Combined Nomenclature commodity code. CBS-IRIS uses this feature to automatically translate company codes into the defined commodity code;
- one of the most difficult data elements for companies proved to be the *statistical value*. For companies who regularly declare the same kind of shipments, CBS-IRIS can calculate the statistical value from the invoice value by making use of user defined algorithms. Different algorithms can be defined for different kinds of shipments.

All these features were developed to *translate* company data into statistical data. The translation rules are prepared by the reporting company.

In the original concept, introduced in 1993, the statistical data had to be put onto diskette – provided free of charge – and sent to Statistics Netherlands. Recent developments, however, have led to a shift towards other forms of data communication, as we shall describe below.

## **5. Certified software by third parties**

Apart from developing CBS-IRIS, Statistics Netherlands also provided the specifications for the trade statistics data to software houses and to companies who wanted to develop

their own custom-made software. These specifications contained the prescribed record layout for the data as well as rules for validity and consistency checks.

Software developers who had built INTRASTAT tools and companies developing custom-made software for their own use were invited to send test data sets to Statistics Netherlands. Once the format and the validity of the data had been checked, companies were given permission to use their own software to send in INTRASTAT returns. Software companies meeting with the requirements laid down in the specifications – approximately 50 in 1994 – were granted a conformity label. In the statistical process the data received from these systems were treated identically with the CBS-IRIS data.

## **6. Introduction and reception**

The introduction of CBS-IRIS coincided with the introduction of the INTRASTAT obligation. As the obligation to report their intra-EU to Statistics Netherlands was new to many companies an extensive instruction campaign was organised for them, often jointly with local Chambers of Commerce. In addition to information on INTRASTAT, every session also included a demonstration of the CBS-IRIS tool.

At the end of 1992 and the beginning of 1993 some 12,000 CBS-IRIS packages were handed out or sent to trading companies and administrative offices. The response was overwhelming.

### *Problems*

Companies contacted Statistics Netherlands with numerous questions about INTRASTAT, and also about the CBS-IRIS-package and its documentation, which was apparently not self-explanatory enough. In addition to these queries other aspects did not exactly facilitate the introduction of CBS-IRIS. The computer equipment used by the companies was much more diverse than expected: equipment and operating systems that we believed to be obsolete were still operational at some companies. And it should also be stated that the first release of CBS-IRIS did contain some bugs, some of them even major bugs.

These problems gave rise to very intensive telephone contact between the companies and Statistics Netherlands, sometimes to such extent that the lines became overloaded and telephone communication was no longer possible.

### *Solutions*

The problems encountered during the introduction of INTRASTAT and CBS-IRIS program all called for specific measures:

- Having extended the number of available telephone lines with the aid of the telecommunications company, a help desk was installed. In the first months of 1993 this desk was manned by up to 25 people, some of whom had to be recruited from other departments within Statistics Netherlands as this was too great a demand for the foreign trade statistics department alone. Crash courses were organised for those not familiar with INTRASTAT and CBS-IRIS.
- Fieldworkers from the foreign trade statistics department started an intensive campaign, visiting companies, assisting them with the installation of the CBS-IRIS package and giving more specific information on INTRASTAT. The capacity of the group of fieldworkers was enlarged considerably.
- A comprehensive manual was compiled for CBS-IRIS, containing detailed information on the data elements required and thus answering many INTRASTAT related questions as well. The use of the extra features in CBS-IRIS was also explained in detail.
- A number of bug releases of the software were sent to the companies experiencing problems that could not be solved otherwise.

Naturally, the introduction of CBS-IRIS simultaneously with INTRASTAT required some extra effort from Statistics Netherlands. However, a few months after the introduction the situation gradually stabilised as companies got used to the INTRASTAT system and the CBS-IRIS software.

The help desk remained essential in the support of CBS-IRIS and thus in the acceptance of CBS-IRIS as a standard tool for the INTRASTAT returns, but it was possible to reduce the help desk staff to ten. By January 1994 9,000 CBS-IRIS users were known to Statistics Netherlands, including the third party providers, sending in INTRASTAT returns for up to 14,000 companies trading abroad.

#### *Data communication*

Companies called the help desk not only with questions about CBS-IRIS or INTRASTAT, but also with suggestions to improve the functionality of the program. One of these suggestions was a data communication option alongside the possibility of transferring the data to diskette. This prompted Statistics Netherlands to develop a data communication module and implement it in the CBS-IRIS program, although before this option could be made operational, security precautions had to be taken.

First of all the data sent by the companies should be protected against unauthorised access. Furthermore Statistics Netherlands has to be protected against infected files on its Wide Area Network. To meet these two demands without making data communication overly complicated, the following solutions were chosen. The data files sent by CBS-IRIS are encrypted and compressed. The upload server at Statistics Netherlands is a stand alone computer and there is a mail lock between this server and the Wide Area Network in the shape of a computer that connects to the server to retrieve

uploaded files. It then disconnects, and the uploaded files are decompressed and decrypted, checked for viruses and for recognisable format. If files are suspect or if their format is not recognised, they are rejected and deleted. The mail lock subsequently connects to the Wide Area Network and the approved files are imported there. This method of data communication is essentially one way. The sender only receives acknowledgement of receipt.

Up to now this concept has not posed any serious security problems for Statistics Netherlands. The data communication option was introduced in CBS-IRIS at the end of 1993, for a restricted number of users at first. After the first positive experiences the communication module was built into the subsequent CBS-IRIS release. Although the data communication option was not widely advertised, the number of enterprises using the option increased steadily from 200 in the spring of 1994 to over 2,000 users in January 1997.

#### *Release policy*

After the initial period, when some bug releases were made, Statistics Netherlands followed a strict release policy. The CBS-IRIS program is updated once a year, and the new version is sent to all users in the last week of December and the first week of January. The new releases incorporate improvements resulting from users' suggestions and the officially revised Combined Nomenclature and the Country Nomenclature code lists.

## **7. Consequences for the organisation**

The introduction of EDI/EDC had a considerable impact on the organisation of the foreign trade statistics department. As described above the effects of the various projects – the introduction of INTRASTAT, the introduction of CBS-IRIS and the customs automation project Sagitta – cannot be seen separately. The most obvious change in the organisation was the reduction of required data entry capacity. In January 1993 the data entry staff consisted of 70 full-time employees. As CBS-IRIS and Sagitta came into effect, and information received on paper decreased, this number dropped to 50 full-time equivalents in January 1994, and even further to 12 in 1996.

However, the implementation of EDI/EDC not only led to job losses, new jobs were created too. The essential role of the help desk in the introduction of CBS-IRIS has already been mentioned above. This was set up with a staff of 25 at the start of the project, although the number was reduced to ten in 1994 and even further to five when things had settled down at the end of 1995.



In addition to the installation of the help desk, the handling of the influx of diskettes involved new tasks in the organisation. An extra capacity of two full-time equivalents was needed to dispatch the diskettes to CBS-IRIS users and to process the returned data diskettes. Another four people were needed as troubleshooters, not only for problems with CBS-IRIS diskettes, but also to check products of software companies and assign them conformity labels, and to process tapes and diskettes from companies using custom-made software. With the increase of direct data communication it now seems that this work will require less capacity.

All in all, it is obvious that with the switch from a stream of information on paper to the use of EDI/EDC some new fields of attention have been developed. Although this has led to a considerable reduction in the required data entry capacity, it has also led to the creation of a number of jobs requiring higher qualifications.

## **8. Consequences for the statistical process**

Naturally, the fact that data are checked while being processed at the source – the reporting companies – improves the quality of the individual data received by the foreign trade statistics department. The receipt of the data in a machine readable form makes the process of checking for consistency less tedious, both for the intra-EU trade statistics and for the extra trade statistics. Where less attention has to be paid to detailed data checks, more can be focused on the statistical process. Where the introduction of the INTRASTAT system introduced aspects such as partial response and non-response to the international trade statistics at least for the intra trade statistics, this changed the process from an administrative process to a real statistical process.

EDI/EDC added to this change a higher reliability of the detailed data, and a decrease of time needed for the basic throughput, thereby creating the possibility to spend more time and capacity on the real statistical process.

Of course, some critical remarks can be made on the effects of the use of CBS-IRIS. Although the introduction of the algorithms to calculate statistical value makes life easier for the companies, it also involves the risk that these companies will not evaluate them from time to time. In the long run this may have a negative effect on the correctness of the figures for statistical value. Similarly, if the default values are not updated from time to time, it is possible that incorrect data will be introduced. Research into the effects on statistical value is planned for the second half of 1997.

## 9. Evaluation

With hindsight, we could ask ourselves whether the introduction of CBS-IRIS on such a large scale was a wise thing to do. Certainly Statistics Netherlands encountered considerable problems through the simultaneous introduction of INTRASTAT and sending out 12,000 copies of CBS-IRIS and, looking back, it took a lot of extra effort. But in the end we were able to solve most problems acceptably for all parties concerned. The advantage of the CBS-IRIS introduction procedure was that it was accepted as a suitable tool to handle INTRASTAT returns by most companies. So the answer to the question at the beginning of this paragraph is: yes.

Although the assimilation threshold was raised twice in the past years, this did not lead to a great reduction in the number of CBS-IRIS users. Every year somewhere between 850 and 1,000 companies ask for evaluation copies of CBS-IRIS, and most of these become CBS-IRIS users. Even after the two assimilation threshold increases, the number of enterprises and administrative offices using CBS-IRIS has remained fairly constant at approximately 9,000.

Reactions from companies are mostly positive, although with the evolution in PC operating systems we are receiving more and more requests for a Windows version of the software, which – incidentally – Statistics Netherlands intends to introduce in the course of 1997.

In the long run it is considered a disadvantage that the data required for the intra-EU trade statistics connect so poorly with the data available from the companies' own administrative systems. If we could achieve a better link between these two data sets, a more advanced method of EDI would be possible, reducing the administrative burden for the reporting companies even further.

CBS-IRIS was one of Statistics Netherlands' first EDC projects. Its success has certainly made the bureau more aware of the possibilities of EDI and EDC, and in this way CBS-IRIS' role in the introduction of EDC/EDI in data collection has been an important one.

## References

European Union (1991-1). Council Regulation (EEC) No. 3330/91 of 7 November 1991 on the statistics relating to the trading of goods between Member States.

European Union (1991-2). Council regulation (EEC) No. 3330/91.

European Union (1992-1). Commission regulation (EEC) No. 2256/92 of 31 July 1992 on statistical thresholds for the statistics on trade between Member States .

European Union (1992-2). Commission Regulation (EEC) No. 3046/92 of 22 October 1992 laying down provisions implementing and amending Council Regulation (EEC) No. 3330/91 on the statistics relating to the trading of goods between Member States.

European Union (1992-3). Explanatory notes to the Intrastat forms referred to in Article 2 of Commission Regulation (EEC) No 3590/92.

Statistics Netherlands (1994). Blaise – a survey processing system.

# The TELER-EDISENT project

*Gerrit W. de Bolster and Kees J. Metz*

*The TELER project is a joint operation of several European statistical organisations. It focuses on the collection of statistical information using EDI from enterprises. Several possibilities have been researched, of which EDISENT in particular is discussed here. EDISENT is a module, a piece of computer software which translates business information into information required by the statistical institute. The TELER-EDISENT project is a form of primary EDI where both the respondent and the statistical institute are responsible for the translation from business to statistical information.*

## 1. Introduction

The use of EDI techniques has been under discussion for quite some time. One example is the discussion by the SERT Ad Hoc Group, convened by Eurostat in Luxembourg twice a year (SERT is a French acronym for Business Statistics and Telematic Networks). Here, the idea was launched to join forces in a multinational project on the use of EDI in business statistics and to apply for subsidies from the 4th Framework Programme for Research, Technical Development and Demonstration, launched by the European Commission. A proposal for a project named TELER (TELEmatics for Enterprise Reporting) was submitted under the Telematics for Administration Programme in 1995 and was accepted by DG XIII for funding from the 4th Framework Programme. Actual work on the TELER project started in mid January 1996, and according to the project programme should last to March 1999.

The TELER project aims to demonstrate the feasibility of a common model and related standards for the exchange of data between enterprises and data collectors, especially national statistical institutes (NSI's). Both enterprises and NSI's will benefit from the implementation of this approach as it offers:

- simplification, e.g. by merging several statistical surveys into one electronic combi-questionnaire;
- common standards, by adopting one model, whatever the size of the enterprise, sector of industry, country, accounting practices, etc.;
- gains of workload, as the technologies used will (i) allow tapping of electronic data provided by the accounting packages or invoices systems used by the enterprise (which saves data entry), (ii) help transcode these data into national and international

classifications, (iii) provide transmission facilities to send the data with the appropriate structure.

## 2. Scope

TELER will build data classifications, data models, interfaces and other software valid for all Community members. The demonstrator to be built is based on:

1. a repository or model of data necessary for the data collectors, called BISE;
2. software modules with facilities for format translation, BISE management, communication, etc. called SISE<sup>1)</sup>.

The TELER project involves three different kinds of data collectors: NSI's, national professional associations in the steel industry (EUROFER) and accountants (EDIFICAS). They will implement and test the databases (BISE) and applications (SISE) on nine sites in eight countries. Three types of data exchange will be tested:

- variant 1 (EDISENT<sup>2)</sup> sub-project): company to NSI, the BISE being at the NSI (seven sites in seven countries: the Netherlands, Sweden, Germany, Italy, Finland, Portugal, Spain);
- variant 2 (EDIFICAS sub-project): company to accountant and accountant to NSI, the BISE being at the accountant (five sites in five countries: France and, carried out by the NSI under the supervision of EDIFICAS: the Netherlands, Sweden, Germany, Italy);
- variant 3 (EUROFER sub-project): company to national professional association (NPA) and NPA to Eurostat, the BISE being at the NPA (one site in France).

In variant 1 information for business surveys (including short term indicators) will be collected at the enterprises in an automated way. This information consists of financial data as well as production data. At every site ten to twenty manufacturing companies will be involved in the trial. The number of different variables to be collected during these trials varies from country to country:

<i>Country</i>	<i>Number of variables</i>
Germany	55
Finland	316
Italy	456
Netherlands	242
Portugal	155
Spain	262
Sweden	295

In variant 2 the information will be restricted to financial data. One or two accounting firms will be involved at all five sites. In variant 3 the focus is on 'delivery data', which are very difficult to manage, especially in the case of triangular operations.

This article focuses on variant 1, for which the NSI's are responsible. Trials will be carried out using demonstration software called EDISENT. This EDISENT module will also be used in the trials of variant 2 in The Netherlands, Sweden, Germany and Italy, under the responsibility of EDIFICAS.

### 3. The TELER consortium

The TELER consortium consists of:

- CESIA (a French consulting firm) acting as a consultant for the organisation and as a project manager (it is not a software house);
- various data collectors, such as the NSI's in seven countries and accountants and professional organisations acting on behalf of enterprises, who are the users and conduct the project;
- a few IT companies who help them implement the functions and standards agreed upon in early stages.

Moreover the User Group validates each individual step.

<b>Partners</b>	<b>Corresponding associate partners</b>
CESIA (France)	
Statistics Netherlands	– INE Portugal – CAP Gemini
Statistics Sweden	– Statistics Finland
ISTAT Italy	
Statistisches Bundesamt-Germany Nordrhein/Westfalen	– Landesamt für Datenverarbeitung und Statistik – Landesamt für Datenverarbeitung und Statistik Brandenburg
Eurofer Europe (with its members: ISSB-GB, CPS-FR, WVS-DE)	
Edificas Europe	– Edificas France – Datacare

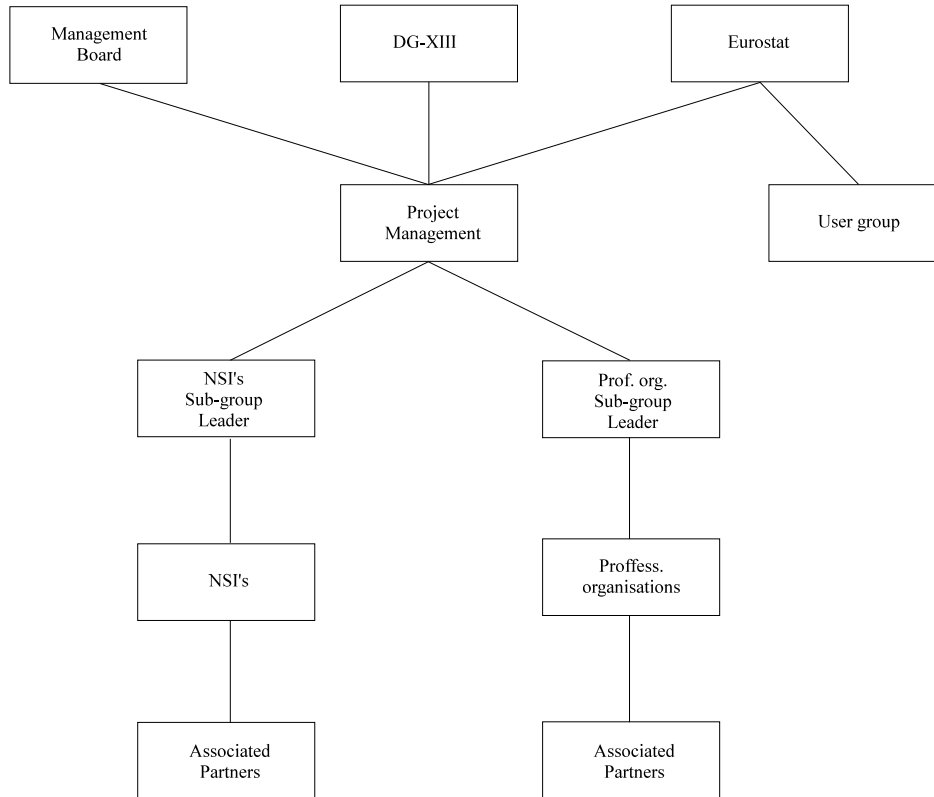
INE, the Spanish NSI, is participating in the NSI group on a voluntary basis and is not an official member of the consortium.

*Overall organisation*

The overall organisation of the project is as follows:

*Overall organisation*

The overall organisation of the project is as follows:



The consortium is made of two sub-groups, one consisting of NSI's and one of professional organisations. The project leader is the leader of the latter. An NSI (Statistics Netherlands) is the leader of the NSI sub-group.

*User group*

The User Group is linked to the SERT Ad Hoc Group and is thus composed of NSI's from the European Union not already within the consortium and a number of professional organisations. EFTA countries and others might send representatives. Additional data collectors will be sought (chambers of commerce, other government

institutions and more professional organisations, etc.). The participation of as many NSI's as possible is important for two reasons:

- so that the TELER developments can be adjusted to the various national situations as far as classifications, accounting practices, and norms are concerned, and can take into account the heterogeneity of enterprises across countries and across sectors of industry;
- the NSI's wish to keep control of their own statistics and statistical methods, based on the principle of subsidiarity.

The User Group meetings are convened in cooperation with EUROSTAT and will link up with similar statistics related projects.

#### 4. EDISENT

Up to now, NSI's have often sent companies or organisations separate questionnaires for every statistical survey conducted. In order to complete these questionnaires, the data providers must draw information from several separate administrative accounts, and then combine and/or recalculate this information to bring it into the format requested by the data collector. They then have to repeat these actions every time they receive the questionnaire, yearly, quarterly or even monthly.

Sometimes, the same or similar questions are included in several questionnaires, causing unnecessary overlap. Implementation of modern information technology can help to change this. Instead of starting from the NSI's point of view, from now on the automated accounts at the enterprises should be considered as the starting point for data collection. Instead of a questionnaire for each statistical survey, a questionnaire for each automated account will be developed. This will be an electronic questionnaire, combining questions for several separate statistical surveys into one so-called combi-questionnaire. In TELER the following number of combi-questionnaires will be used in the trials:

<i>Country</i>	<i>Number of combi-questionnaires</i>
Germany	5
Finland	2
Italy	3
Netherlands	5
Portugal	2
Spain	2
Sweden	1



A precondition for this approach is that concepts and definitions used by the NSI as data collector should be attuned to those used by the data provider. Sometimes the information required at the NSI can be drawn from the accounts at the enterprise or institute directly, by introducing provisions in the software that the data provider uses, enabling him to supply the requested information just by pressing a extra key. In many cases, however, the companies and organisations concerned use varying concepts and definitions and require a very flexible tool to be able to provide the requested data in an automated fashion. For this purpose, the EDISENT-module will be developed for TELER. This module, a tuneable translator, must allow for the automated provision of the majority of the data to be collected, in a format that can be used by the NSI as data collector, thereby reducing the effort required on the part of the data provider.

It should be stressed that it is the data provider who decides *whether* and, if so, *when* and under which conditions, the data from his automated accounts will be made available in this way. The 'EDIfication' process works both ways, as it will lead to a faster method of data collection. Due to EDI, the time needed for data processing (e.g. for validation and editing) can be reduced considerably. It is also anticipated that this will also improve the quality of the collected data, in comparison with data supplied via the manual actions required to complete the current questionnaires.

*The proposed method of the EDISENT module*

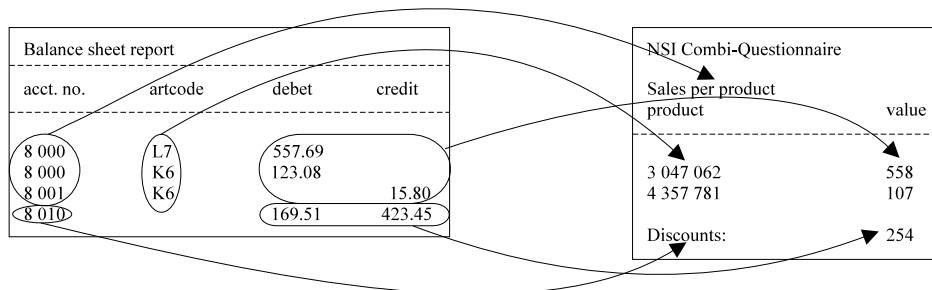
The contents of the combi-questionnaire are dictated by what is available in the financial accounts. As regulated as our society may be, the financial accounts may diverge strongly in organisation and in the concepts they use. In the first place this means that we will have to adapt our questions to the possibilities of the companies' automated systems, with the possible implication of more statistical work for the NSI to reach the same output. If more information is required, it will probably have to be explicitly requested and will result in data entry. In the second place the diversity of respondents means that a unique translation scheme will have to be set up and maintained for each respondent.

Financial accounts also differ in their technical layout as they are based on a variety of bookkeeping software systems. There is no standard record for information to be selected electronically from the software and it is not expected that it will be possible to define one in the near future.

As the main goal of EDI is to reduce the response burden, it was decided that the amount of data entry was to be minimised. Creating the automated link we were looking for required quite some ingenuity. It is done by using the reports or printouts of the software system. Instead of being printed, they are sent to a print file to be read by the translator, the main part of the software module that is currently being developed as part of EDISENT to run on the respondents' computer. The layout of the reports – and thus of the print files – is fairly stable. The respondent communicates this layout to the

translator, defining rows and columns within the report. Subsequently he 'instructs' the translator how to manipulate the rows and columns in order to convert the information in the report to the statistical information asked for by the combined questionnaire. The resulting records are transmitted to the NSI.

*The translator*



We see then the two parts of the translation scheme. The first part lays down the lay-out of the printfiles to make the technical transformation. The second part defines the conceptual transformation of the information in the printfile to the statistical information asked for on the combined questionnaire.

The final question is who will make this translation scheme. As one of the principles of EDISENT is that *the respondent translates*, it is the respondent who has to set up the translation scheme. Although this naturally makes it less respondent friendly, it has proven to be impossible to set up the translation schemes at the NSI's. Clearly, this will not be an easy task for the respondent and will require a well-staffed support desk and a fairly large fieldwork service on the part of the NSI. Even then EDISENT will not achieve the ultimate user-friendliness of EDI. So ideally it must be made easier to tune the EDISENT module, which could be done by persuading both parties, respondents and data collectors, to use the same accounting concepts. To this end the TELER project has been decided to extend the EDISENT module with a repository based on a structure proposed by EDIFICAS. If this structure, also called coding language, evolves into a common structure for accounting systems accepted by all parties, it may even make translating, and thus tuning, obsolete in the future. In the meantime respondents and data collectors can use this repository as a link between their conceptually different worlds.

We expect the translation scheme to be fairly stable or, in other words, that technical and conceptual changes will not be too frequent. For a second delivery of data the translator can use the already available translation scheme to produce the statistical information. Answering the combi-questionnaire then becomes a matter of minutes instead of hours and can be handled by a less qualified employee. This is what makes the concept attractive and the initial investment worthwhile to the respondent.

## **5. The process at the NSI**

NSI's using the EDISENT concept will no longer receive separate statements from respondents for separate surveys, but instead statements with information for a number of surveys. These statements must therefore be stored centrally so that the statistical departments can select input for their particular surveys. This central collection and storage of data – the input database – will change the statistical process inside the NSI. Up to now a statistical department could change the contents of its questionnaire annually. Because of the once only tuning in the EDISENT module it is of great importance to stabilise the contents of the combi-questionnaires. Every change will not only cause the need for an update of the installed EDISENT module but also the adjustment of the tuning, resulting in a burden for the respondent. Although this would seem to limit the freedom of the statistical department, the opposite is true! The input database contains the data collected for all surveys of enterprises. At the NSI this means that almost all the administrative information available at enterprises is collected. A statistical department can shop around in the input data base for more information without bothering the enterprises.

We still have to keep in mind that the information in the input data base is divided according to the different respondent sources, and the statistical departments have to combine these pieces of information into the information about the statistical units they need for their input. Moreover, the collected information consists of administrative, not statistical, data items. As mentioned before, the translation will have to be done inside the NSI.

The communication between the statistical departments and the respondents must also be coordinated. Naturally, it is unacceptable for several statisticians to contact the same respondent with similar queries about his statements! Although this way of collecting data will cause more – and more complex – work for the NSI, it creates opportunities for the statisticians to combine more information of higher quality. We must not forget that the main reason for applying EDI for data collection is to lower the administrative burden for the respondent!

## **6. Progress**

The first phase of the project, from January 1996 to September 1996, focused on the current situation. For variant 1 an inventory was made on the data flows from enterprises to the NSI's, accounting practices were investigated in several countries and a small survey was held to determine the use of standard software and the types of operating systems for business administrations. After an internal review at the end of 1996 the project programme was adjusted. In the first half of 1997 a common conceptual data

model for all three variants was developed. An international prototype of the EDISENT module was built for variant 1. Shortly after that the global specifications of the software used in each of the three variants was produced. At the time of writing the sub-groups are working on the detailed specifications of the software and, simultaneously, a start has been made with building this software.

## Notes

- 1) BISE and SISE are acronyms for Base d'Information Statistiques d'Entreprise (Statistical Enterprise Information Basis) and Système d'Information Statistique d'Entreprises (Statistical Enterprise Information System). The concepts of BISE and SISE were introduced in the SERT study 01 that the French consulting firm CESIA conducted on behalf of Eurostat.
- 2) EDISENT is an acronym for EDI between Statistics and ENTERprises. The name is used for both the concept and the software module that will be used to transform the data in the enterprises' automated accounts into data collected by the NSI's and others.

# Electronic supply of data for labour statistics

*Freek Arnoldus*

*Payroll administration in the Netherlands is strongly regulated by fiscal law and social security regulations. Traditionally Dutch labour statistics have used several 'paper' surveys with a relatively high burden for the respondent. In the context of the EDIfication of the collection process Statistics Netherlands has defined standard deliveries from the existing payroll administrations, making it possible for the owners and developers of these systems to develop – in consultation with Statistics Netherlands – Statistics Netherlands functions to be incorporated in the system. This then replaces the paper questionnaires by a monthly or quarterly delivery of a detailed file from the payroll administration.*

*This project is a successful example of primary EDI. The translation of the concepts is done by the respondents, although they can consult Statistics Netherlands for the definition of the translation scheme, and the requested concepts were chosen close to those directly available from the payroll administration. In October 1997 the data from over 14.5 thousand enterprises, describing 2.5 million employees, were supplied using this new electronic method.*

## **1. Labour statistics**

Labour statistics, in this case those on employment and wages in particular, are compiled on the basis of data from several surveys among companies and institutions (from here on we shall refer to these together as *companies*). These comprise two surveys on a quarterly basis, three annual surveys and one that is conducted every four years. Apart from aggregated data like wage sum and number of employees, both the quarterly and the annual surveys also ask for data on individual employees like age, hours worked and wage rate.

The population of all the surveys are all companies with employees, regardless of economic activity, with the exception of one survey that is held annually among local government institutions only. Companies are obliged by law to respond.

The surveys are sample-based, with the sample fraction depending on the number of employees of the enterprise: it rises with the number of employees. In most cases the fraction equals 1 for companies with 100 or more employees. For individual employee

data, a sample from all employees on the pay-roll of the companies is used. As the total number of employees grows, the fraction of the sample of employees becomes smaller.

If these data were collected by questionnaire alone – which has no longer been the case for some time now – the number of questionnaires would be over 250,000 on a yearly basis, sent out to 90,000 enterprises. In this hypothetical situation, information would be collected on approximately 400,000 individual employees annually and 40,000 employees every quarter.

## **2. Payroll administrations**

The law covering wage administration is rather complicated in the Netherlands. For instance, personal circumstances of individual employees and measures to stimulate the employment of specific groups in the labour force define the level of several wage levies and allowances. Furthermore, the execution of the different types of social insurance differ by branch. This all leads to strong differences in payroll administrations.

The organisation of salary administration has largely been determined by this complicated system of laws. Establishing the level of the taxes and the contributions for the different forms of social security and other contractual obligations, and collecting and transferring the money involved are all standard activities of the wage administration. These activities involve an intensive flow of messages between the employing companies and the social security institutions. Not complying at all, complying too late or not fully complying with the demands of these organisations is punished with severe fines. As a lack of appropriate knowledge of social security laws can therefore be very expensive for the enterprise involved, many – mostly smaller – enterprises have placed their payroll administration in the hands of specialised bookkeeping agencies or accountants. Larger enterprises often transfer their salary administration – including the distribution of the resulting data – to computer service bureaus. It is estimated that the wage administration of about seventy percent of all employees in the Netherlands is done through one such bureau. For most of the remaining employees the administration is done using standard software. Only a limited number of enterprises, mostly large companies, use a tailor made system for their wage administration.

As the exchange of data between the companies and the executive institutions is done mainly electronically and is a standard facility of a payroll system, it seemed increasingly likely that Statistics Netherlands would remain the only organisation requiring data to be supplied on paper questionnaires.

### 3. Electronic data supply to Statistics Netherlands

At the request of some of their larger clients, in the early eighties some computer bureaus made provisions in their payroll systems for the wage data requested by Statistics Netherlands. Most of these provisions related to surveys which the companies experienced as the most labour intensive. Some large companies with their own wage administration systems also developed such facilities.

Although Statistics Netherlands provided record descriptions, the respondents used a diversity of technical and conceptual formats. The number of respondents making provisions in their systems for the less labour intensive surveys remained limited.

At the end of 1993 Statistics Netherlands started looking into the possibilities of the automated payroll systems for the statistics on wages and employment. The aim of the study was to develop a method of data collection which minimised the efforts of the responding companies, and the main focus was on the possible role of information technology. For reasons of effectiveness and efficiency the study started with the payroll systems of the computer service bureaus. Some twenty of these bureaus account for the administration of 70% of all wage transactions in the Netherlands.

This initiative by Statistics Netherlands was received positively by the service bureaus. This was not really a surprise as the service bureaus tried to present themselves as the central points of collection and distribution of payroll information. Later on in the study developers of standard and of tailor made software were involved. They too reacted favourably.

The study revealed the following bottlenecks in the new method of electronic data collection:

#### *The diversity of wage statistics*

The various wage statistics were developed at a time when wage administration was mainly done by hand. To minimise the work involved for Statistics Netherlands and for the companies concerned only the data actually needed for particular statistics at that moment in time were requested. This led to differences in methods and questions between surveys. The different sample fractions of companies and employees within the companies, is a case in point. The fact that coding for the same groups of employees is not always the same is another. Therefore several service bureaus regarded the development and maintenance of a special function in their software for the electronic supply of data for consecutive wage statistics as too expensive.

*The requested data are not always part of the standard data*

The organisation of the wage administration is largely determined by the so-called *gross – net trajectory*, which starts from the gross wage agreed upon (before taxes and social security contributions) and ends with the net wage as paid to the employee. For reasons of efficiency, large salary systems in particular, like those of the computer service bureaus, record only the information they need for periodical wage computations. Items which are indispensable for the wage computations in one branch may be of no interest to another branch. The number of days leave, paid sick leave and some employee characteristics, for example, are not relevant for the wage administrators, but are very interesting for Statistics Netherlands. These data are usually present in the decentralised employee administrations at the companies themselves.

Similar problems also occur with regard to wage concepts used only by Statistics Netherlands, which are not included in the gross-net trajectory. Frequent changes in the law render the wage concepts used for taxes and social security unsuitable for statistical purposes. Therefore, statistical wage concepts have been defined that are not or only slightly affected by administrative measures. However, as these concepts do not play a role in the computations of the wages, the necessity of incorporating them in the payroll administration is very small.

*The moment at which data have to be delivered*

December, January and February are the busiest months for salary administrations. In addition to several departments in the companies concerned, the financial accounts department for example, several external organisations also have to be provided with data on the previous year. The most labour-intensive Statistics Netherlands survey from the companies' point of view, the annual wages survey, is also conducted in January. Companies do not conceal the fact that although they are obliged by law to participate in this survey, Statistics Netherlands features very low on their list of priorities at this time of year.

#### **4. The new method of electronic data collection**

After much consultation both inside Statistics Netherlands and with several specialists in the field like wage administrators and service bureau staff, a new method to supply data was developed in 1994. It makes it possible to derive company data directly from the wage system and transmit them to Statistics Netherlands on magnetic tape or diskette. Both the fact that the data can be derived directly and the fact that the service bureaus can deliver straight to Statistics Netherlands contribute to a large extent to reducing the administrative burden.



The most important characteristics of this method, which serves as an alternative for the consecutive paper questionnaires, are:

*One set of specifications for several statistics*

The method in fact involves the union of the concepts collected through different questionnaires. This is also true regarding the different methods of data collection. In other words: various surveys have merged into one, with the advantage that only one computer program has to be developed and maintained to supply the data for several statistics.

*One source: the wage administration*

The starting point for the data specifications was the data that are usually included in automated wage administrations. So information not usually included in the automated wage administration but held instead in the decentralised employee administrations are not included. The missing data are collected from other sources.

Data that cannot be tapped directly from the wage administration data but have to be derived from the available data, are compiled at the source, not at Statistics Netherlands. Because of the diversity of the wage concepts among the various branches the number of concepts necessary would be too high. Instead, for each wage administration, a procedure has been defined – in close collaboration with Statistics Netherlands – to translate the available wage concepts to the required statistical ones. Furthermore, the definitions of the statistical concepts make reference to the 'hard' concepts used by the internal revenue services and the social security institutions.

As in most other larger EDI projects the information flow to Statistics Netherlands is redesigned. All in all the amount of information collected for the labour statistics will hardly diminish. Most of the paper surveys mentioned here will eventually disappear. Part of that information will be collected by EDI with the wage administrations, the main subject of this article. After translation, the concepts will in effect cover the same statistical information as before. This means that the once-only definition of translation procedures for these concepts will require a lot of time and effort. The remaining information required by Statistics Netherlands is obtained from other sources with other methods.

*Connection to existing standards*

For file structure, record layout, coding, and data transmission, Statistics Netherlands tried as far as possible to conform to existing and widely used standards and procedures in computerised wage administrations. To a large extent use was made of the standard already in use by the internal revenue for the yearly returns by all companies on their employees. This standard does not discriminate by branch of economic activity or by any other criterion and had been in use for some time.

#### *Comprehensive delivery without selections*

For many salary systems sampling is an inappropriate activity, resulting in extra costs compared with the supply of comprehensive files. The same is true for selections of records used only by Statistics Netherlands. In the new method all records present in the system at a certain moment in time are transmitted to Statistics Netherlands, so that the selection of particular groups of employees and former employees has become an activity in the statistical part of the chain. This means that particular changes in legislation will no longer affect the respondent, but will lead to changes in the statistical selection procedure.

#### *Deliveries by third parties*

The new method makes it possible for others than the company itself to ensure the delivery of data to Statistics Netherlands: for example accountants' offices and computer service bureaus. Here too an existing standard for third party data supply has been introduced.

#### *Choice of frequency of delivering data*

The data are required by Statistics Netherlands – and therefore by law – on a quarterly basis. However, if, for reasons of efficiency, a company would rather supply the data every month or every four weeks, they can. In many cases the Statistics Netherlands function is linked to existing system routines. Often the wage account files for individual employees are produced in the same run as the files to be used by Statistics Netherlands.

### **5. Implementation of the new method**

By the end of 1994 Statistics Netherlands had sent the manual on the new electronic method to all known computer service bureaus and salary administration software developers, followed up in most cases by a further personal explanation. In addition to the technical manual, a brochure for the users of wage administration systems was composed describing how the method works. In the first months of 1995 this brochure was sent to about 35,000 companies. Other channels were also used to publicise the new method: Statistics Netherlands was represented at the annual fair on wage data processing; an extensive article on the new method was written for a professional journal for wage administrators; presentations were held at meetings of wage administrators; each letter accompanying paper questionnaires pointed out that there was an easier way to deliver data to Statistics Netherlands. Providers of salary administration services and software also informed their customers that, in consultation with Statistics Netherlands, they had included the Statistics Netherlands function in their pay roll system.

Statistics Netherlands was closely involved in the actual development of its specific function in the wage administration systems. This involved further explanation of the method, commenting on functional designs and, most especially, testing the function.

Without doubt the strong competition between suppliers of computer services and between software developers contributed to the fact that within a relatively short period – three years – these companies were in a position to inform their clients that the Statistics Netherlands function had been or was about to be incorporated in their systems. As is the case for data exchange with other institutions, the clients are charged for the Statistics Netherlands function. By the beginning of 1997 the function was included in the salary administration systems of all the major computer service bureaus and many of the larger accountants offices. Furthermore several software packages included the function and for a number of other packages development was well under way. The tailor-made systems of a number of large concerns also allowed for the data supply to Statistics Netherlands using the new method. At the beginning of 1997 over sixty administration systems had incorporated the new Statistics Netherlands method, and by October of that year the data of over 14,500 companies describing 2.5 million employees were transmitted to Statistics Netherlands by the new electronic method. Half of the data are transmitted monthly or four weekly, the other half quarterly.

Most of the users of the new system are in the government and the care sectors. Furthermore larger companies are better represented than smaller ones. In October 1997, over half of all enterprises with 200 or more employees responded by way of the new electronic method.

In the course of 1997 several wage administration software firms will make the Statistics Netherlands function available to their clients. Based on the positive reactions of companies who have already switched over to the new method, we can certainly expect the share of *electronic* respondents to rise steadily in the future.

## **6. Consequences for the statistical process**

Although Statistics Netherlands received data electronically before 1995, there is an essential difference between the old and the new methods. Before 1995 the method was basically an electronic delivery exactly following the paper questionnaire. The new method on the other hand entails a modified and comprehensive dump from the salary administration system regardless of which survey is being held at the time. This has far-reaching consequences for the statistical process, its organisation, the data processing and the concepts involved.

Originally the various statistics on employment and wages were compiled in separate organisational units responsible for collecting and processing data for their own statistics. As the electronic data deliveries have put an end to these separate data flows, replacing them by just one file, the separate units have been combined into one department.

Another consequence of the new method was a strong increase in the number of data. On the paper questionnaires a company with 100 employees used to deliver yearly some aggregated data and information on 40 individual employees. With the new method the same company now delivers between 400 and 1,300 records, depending on the frequency of supply.

Conceptual differences between paper and electronics, for example the fact that Statistics Netherlands instead of the companies themselves made the selections, also made a new processing system necessary. Developers of the new advanced hardware and software systems gave priority to the function directed at companies and data suppliers, like a control system for the collection process.

Both the mass of data and the conceptual differences between the old and the new methods made the existing data processing system – with checking and correction (editing) as an important part – insufficient. One important difference was the fact that for the paper questionnaires on individual employees, the responding company itself had to select employees, while in the new method Statistics Netherlands does this selection from a comprehensive delivery of data on all individual employees. Therefore a new system had to be built for the selection. The so-called administrative corrections by the companies are another new phenomenon in the electronic data delivery, especially when the Statistics Netherlands function is incorporated in the regular salary administration. An error in the computation of wages in month T is usually corrected in month T+1. But from a statistical point of view, both the data on the wages for T and those for T+1 are distorted. Furthermore, it would be wrong to assume that files supplied by computer service bureaus can only contain systematic errors. Apart from wrong use of the system due to wrong interpretations of the law or insufficient operator skills, companies have ample opportunity to add their own specifications to the of the service bureau's system. Software companies also offer variations by economic branch. All these factors necessitate not only checks on the complete file but also specific checks on the level of the branch and even on the level of an individual responding enterprise.

Statistics Netherlands has now laid the groundwork of the new system to process the data, and will spend the near future on further refinements of the process.

A complicating factor in compiling statistics, in particular when estimating the size of the population, is the fact that many, mostly small, companies still supply their data on paper questionnaires. This means that individual employee data are partly obtained

through sampling and partly by comprehensive observation. This combination has resulted in adaptations in the processing systems, especially for the returns of large concerns where wage administrations are done differently for the different parts of the concern.

The development of the new processing system had the continuation of the regular statistical information as a main priority. The results of the new electronic method, in combination with the collection of data by paper questionnaires, were published recently. Although a certain delay in publication had to be taken into account due to the development of the new systems, it is expected that this delay will be eliminated in the future.

The monthly data of many enterprises are already available for Statistics Netherlands the following month. As the number of companies switching from paper to electronic data supply is expected to rise further, the present time lag will be made up in the near future and publications of the resulting statistics will be able to be released sooner.

# **EFLO: Electronic exchange of local government financial data**

*Fred G.J. Arkesteijn*

*In 1995 Statistics Netherlands started the EFLO project: the electronic exchange of local government financial data. In this project traditional means of data collection, like paper questionnaires and public documents, are replaced by electronic ones: copying the financial administration, or a relevant part of it, onto tape and sending this tape to Statistics Netherlands, where the information is translated into statistical concepts. At present more than 400 municipalities (out of a total 572) have already expressed their willingness to supply Statistics Netherlands with financial data on tape. Two hundred of these already made use of the system in 1995.*

*The EFLO project is an example of primary EDI, where all the translation into statistical concepts is done by Statistics Netherlands, not the respondent.*

## **1. Electronic exchange via the financial package**

The EFLO project is concerned with the financial administration of Dutch local government: municipalities, inter-municipal corporations, provinces and water control authorities.

EFLO takes as its starting point the software packages used by the local authorities for their financial administration. First, we approached the software producers, asking them to incorporate a special provision in the package enabling users to make a suitable copy of the financial records. We then approached the local government bodies concerned – starting with the municipalities and water control authorities, followed by the provinces and larger inter-municipal corporations – and requested them to use this supplementary software. As so many different processing systems are in use, the conversion to electronic exchange will be phased.

The financial packages of five large software houses have already been adapted to the EFLO system. Since February 1995 two Statistics Netherlands staff have been visiting municipal councils to get them to commit themselves to the electronic delivery of financial files. Since January 1997 they have also visited water control authorities. Statistics Netherlands examines tapes of all respondents agreeing to participate in the

project (there may be several respondents per municipality), testing them on readability, identification, completeness and detail. Once the test tape is approved, specific arrangements are made about the actual exchange (type of tape, dates, contacts). The first time the transmission takes place, the data are backed up by paper copies as well.

By the end of May 1997, 404 municipalities and 15 water control authorities had formally agreed to participate. Overall 571 respondents (48% of all respondents) now supply their financial data electronically. This means that the EFLO project is well on target.

**Table 1**  
**Electronic exchange of financial data by local government, 31 May 1997**



	Realisation		Final goal	
	Municipalities	Other local government	Municipalities	Other local government
Councils agreeing to cooperate	404	15	572	250
Respondents agreeing to cooperate	581	15	900	300
Refusals	1	–		
Approved test tapes	571	1	900	300
Procedures confirmed	571	1	900	300

## 2. Statistical concepts

What is transmitted to Statistics Netherlands is a copy of all or part of the financial records of the local government institution involved. As this information is financially, not statistically oriented, it requires translation. Preliminary examination has shown that it is not feasible for the local authorities to do this translation. They experience the allocation and maintenance of statistical codes as a heavy burden because of the size of the financial files, which can contain up to tens of thousands of bookkeeping records. Furthermore, it would be counter-productive, as one of the aims of EDI is to reduce the response burden as much as possible. Also data coding by just one institution (i.e. Statistics Netherlands) will probably guarantee a better quality of statistical information than if it is done at several places. Therefore Statistics Netherlands itself translates the EFLO files into statistical concepts.

The first time this coding is done it will involve a lot of work for Statistics Netherlands as all file entries require a statistical code. But the amount of work involved eases off with each delivery as the coding links are defined and adopted automatically, and only new entries will need to be coded, while existing codes are checked. The comprehensive

**EFLO across the country**

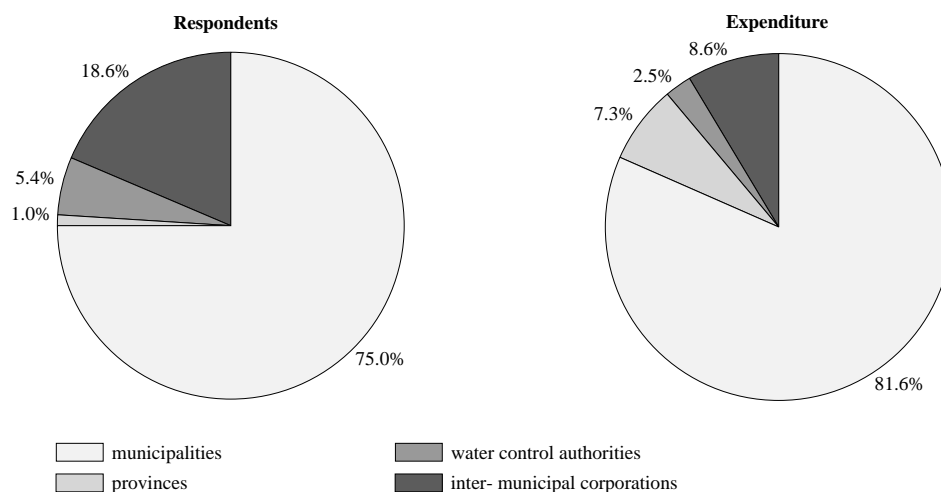
-  municipalities using electronic supply
-  municipalities not yet using electronic supply





files sent to Statistics Netherlands are sometimes so large that processing them presents problems. Therefore a system has been developed to remove superfluous characteristics at an early stage and eliminate unnecessary detail by aggregation, making automatic coding possible.

**Figure 2**  
**Respondents and expenditure by local government - distribution by level of government**



Recently electronically delivered files, for the year 1995, were analysed according to the EFLO method for the first time. Because of the large number of statistical characteristics to be added by the analysts this first time, the analysis was split up into several stages. The first phase entails a general analysis of the files; after that some further processing is done in particular sub-fields, e.g. education, or local construction site operation. Furthermore, more detail is sought for transactions such as investments and intermediate consumption. In the last phase the results of the coding activities are checked for plausibility.

As these phases cannot be carried out in a random order, certain demands are made on the organisation and the availability of the right specialists at the right time. Initially the results were disappointing – in terms of time needed – because of insufficient routine, long waiting times for selections from the files and particularly the sheer complexity of the files. Working with EFLO taught us that using files from financial administrations required extra knowledge as well as extra work compared with the use of paper accounts and questionnaires. The everyday routine of government organisations leads to accounting habits that are sometimes hard to understand, and we often had to use the paper accounts to interpret the files. These accounts give useful extra information, for example on the political backgrounds.

The analyses of the 1996 files will be the ultimate test for EFLO. We hope it will justify the assumption that the analysis of one year can be used to an important degree for the following years. Experience with the limited procedure for quarterly data have indeed shown that the coding rules of earlier quarters can be used in the processing of later files.

### **3. Strict conditions**

Statistics Netherlands strictly adheres to a number of conditions in its electronic exchange of financial data. The financial information is used exclusively to fulfill existing statistical requirements. The files may only be used to compile new statistics if the authorities concerned give their approval. As soon as the statistical processing is finished Statistics Netherlands removes statistically non-relevant information of the authority from the files.

As always Statistics Netherlands guarantees the confidentiality of all data. The national data protection authority has investigated whether the EFLO data exchange conforms with the law on the registration of individual citizens. The EFLO procedure was judged to be permissible under these regulations.

### **4. Consequences for the production process at Statistics Netherlands**

Statistics Netherlands now no longer receives separate financial data on local authorities for the consecutive statistics. It receives all the information for 29 series of statistics (see appendix) in one delivery of data. Therefore a large number of questionnaires will disappear.

After statistical processing the data are stored in the EFLO database. Statistical departments can select the input for their own statistics from this database. They can also search it for other financial information without having to bother the local authority concerned. In this way EFLO ensures the statistical coordination of input data, as statisticians use the same data to produce their statistics. Arrangements have been made for the administration of the EFLO database: use, fixation of datasets, corrections, etc. Furthermore the communication between statistical departments and the local authorities is coordinated, preventing the situation that different departments contact local authorities with more or less the same questions about their statistical returns.

## 5. Investments in the EFLO project

Traditional and electronic data transmission will exist side by side for a considerable time to come – somewhere in the region of six years. As the traditional collection on paper is gradually replaced, Statistics Netherlands estimates that ultimately the procedure will lead to an annual labour reduction of four man-years. To achieve this, seven man-years will have to be invested in the development of the required software within Statistics Netherlands. First-time coding of the EFLO files and the organisation of the project require an extra employment of 2.5 man-years in the first two years of the introduction period and making use of the savings that appear as soon as local authorities started delivering financial data on tape. This requires an extra employment of 0.5 man-years annually for years three to six. Another one and a half man-years is needed annually to visit respondents during the introduction period.

Total investments in the EFLO project also include material expenditure: travel costs, postage, etc. Also the development of EFLO provisions in the financial packages of five software houses cost Statistics Netherlands a few ten thousand guilders.

**Table 2**  
Net investment of the EFLO project, in man-years

	Planning		Realisation
	Total	To 1996	To 1996
Software development	7	3	4.2
Coding and organisation	7	5	4.3
Visiting respondents	9	3	3.6

## 6. Electronic exchange saves time and money

The experiences with the financial packages of the five software houses incorporating EFLO provisions reveal that it takes respondents only 30 minutes on average for each data supply (there are six supplies a year): 15 minutes computer time and 15 minutes of other tasks. The latter is a fraction of the time the respondent would need to complete paper questionnaires. Without exception, respondents supplying their financial data electronically are very positive with this new method.

Of course, the electronic supply of financial information makes the data exchange between local government and Statistics Netherlands a lot more efficient. However, there are also other considerations to change over to electronic delivery:

- Local authorities only need to make one copy of their financial records. In this way Statistics Netherlands no longer needs to ask for supplementary data.
- The administrative burden is reduced, as completion of questionnaires can be omitted. If all local authorities were to supply their financial data on tape, the administrative burden would be reduced by 15,000 hours a year.
- The supplied files contain a greater number of data than the questionnaires, reducing the need for communication between local authorities and Statistics Netherlands about the content of the supplied information.
- Although initially the amount of statistical work for Statistics Netherlands increases due to the bulk of information supplied, each subsequent file decreases the workload and quickens the publication procedure.

## **7. Provisions with other software houses**

At the time of writing, thirteen larger software houses provide the financial administration packages of approximately 95% of the total number of local authorities. As stated above, five of these firms have already extended their software to incorporate EFLO provisions. Consultations have started with the other eight about the possible development of such provisions. The order in which business is done with these companies depends upon the following factors:

- the terms on which the company can develop the required provisions;
- important developments in the financial package at short notice;
- number of users of the financial package and the statistical importance;
- the costs charged by the company.

Statistics Netherlands aims to have incorporated the required provisions in the packages of at least three other software-houses by the end of this year.

When the time has come to replace old software, financial departments in local government have often tended to choose a new financial package from those offered by the thirteen larger software houses. If this trend continues, fewer and fewer local authorities will continue to use their own software or a package of one of the smaller software firms. For any statistically relevant respondents belonging to this group, Statistics Netherlands will deliberate about the way to implement electronic data exchange.

**Table 3**  
**EFLO timetable**

Situation at the end of	Number of software houses with EFLO provision	Number of respondents supplying electronically
1995	3	330
1996	6	600
1997	8	750
1998	10	900
1999	12	1 050
2000	13	1 200

## **8. 100% electronic exchange in 2000**

It is far from efficient for respondents to have to complete dozens of paper questionnaires while they have all the required information in a computer. By the turn of the century the EFLO project aims to receive the financial data of all local government on tape. The introduction period will be phased in this period and for this purpose it is necessary to negotiate with the thirteen software houses. The first phase was completed in May 1997. At the time of writing the EFLO files of the accounts for 1995 of 200 municipalities have been translated to statistical concepts and stored in the EFLO database for the first time. The second phase will be completed in early 1998 when 400 municipalities and 15 water control authorities will have supplied files with their accounts for 1996 and the 1996 EFLO database will have been created. A comparable development is foreseen for subsequent years for respondents with packages from other software companies.

At the moment the EFLO project is still based on data exchange by tape. Although as a medium tape is easy to handle, the use of modems and networks for data communication offers clear advantages in the near future. Within a few years GemNet, a data-communication network for municipalities, will be in use at all municipalities. In May 1997 Statistics Netherlands was also linked up with this datanet and the exchange of financial data via this net will be tested in the very near future. Many municipalities participating in the EFLO project have already expressed their willingness to exchange financial information via GemNet and we expect most local authorities to be transmitting their financial data to us via this network by the turn of the century.

## Appendix

### Surveys included in the EFLO project

1. Progress of ground, water and road construction works
2. Waste collected by municipalities; part B. Expenditures and receipts
3. Liability position of municipalities and provinces
4. Capital expenditure and receipts of municipalities and provinces
5. Statistics of government expenditure on education
6. Expenses and financing of the fire services
7. Expenses and financing of museums
8. Art education\*
9. Art libraries\*
10. Public libraries\*
11. Hospitals\*
12. Mental hospitals\*
13. Nursing homes\*
14. Public health services\*
15. Industrial health services\*
16. Ambulance services\*
17. Pharmacies\*
18. Statistics of municipal budgets
19. Statistics of municipal accounts
20. Statistics of provincial finances (accounts and budgets)
21. Statistics of finances of water control authorities (accounts)
22. Statistics of inter-municipal corporations (accounts)
23. Expenditure on roads and waterways
24. Expenses and financing of legal protection and security
25. Government finances of social welfare
26. Government finances of culture and recreation
27. Expenses and financing of the environment administration
28. Expenses and financing of the landscape administration
29. Administration of water quality; purification of polluted water

\* Only the financial part.

## VIS as secondary EDI source

*Geert Bruinooge, Frans P.M.M. Nijsten, Hen J.M.V. Pustjens and Eric Smeets*

*One of Statistics Netherlands current strategic objectives is to reduce the response burden on the business community caused by statistical surveys by, among other things, making more use of existing central registrations. One of those registrations is the inland revenue corporate tax database VIS (which stands for *Vennootschapsbelasting Informatie Systeem*, which translates as corporate tax information system). VIS contains, among other items, the annual fiscal accounts of Dutch corporate taxpayers.*

*As from statistical year 1995, part of the statistics on finances of enterprises (SFO) is made using information from the VIS instead of survey data. This has cut the response burden caused by the surveys concerned by more than forty percent. This 'VIS to SFO' project is an example of secondary EDI.*

### **1. The corporate tax information system (VIS)**

#### *Population and unit*

The Dutch corporate tax information system (*Vennootschapsbelasting Informatie Systeem* or VIS as we shall call it from here on) contains information on all payers of corporate tax, around 275,000 in number in 1995. These corporate taxpayers, also referred to as VIS units, may be a single company (legal unit), or a combination of several companies in one fiscal unit, which provides a consolidated tax declaration for the entire unit.

#### *Information content*

The information which is important for statistical work relates to tax declaration data, particularly the blocks fiscal taxable capital and fiscal profit calculation. Although a few categories of corporate taxpayers – to wit those registered on the stock exchange and financial institutions – are exempted from the declaration of fiscal taxable capital and profit calculation, they are recorded in the VIS. In practice, however, these exempted taxpayers do provide the relevant blocks together with the rest of their declarations.

#### *Timeliness*

Twelve months after review year almost 50% of the VIS has been filled. In general the largest companies are the last in.

## **2. The statistics on finances of enterprises (SFO) before 1995**

### *Population and unit*

The target population of the statistics on finances of enterprises (*statistiek financiën ondernemingen*, or *SFO*) comprises all corporate sector non-financial enterprises in the Netherlands; statistical information is derived from companies with a legal form (private limited companies, private companies and cooperative societies). In 1994, there were roughly 100,000 of these non-financial enterprises. These enterprises can either be a single company (legal unit), or a conglomerate of one Dutch parent company with several subsidiaries. Such conglomerates provide consolidated figures.

Enterprises with total assets of more than 10 million guilders are all included in the survey, while companies with total assets of under 10 million guilders are estimated on the basis of a relatively small sample survey. Somewhere in the region of ten thousand enterprises receive an annual questionnaire for the SFO, and some 25,000 hours of work are involved in completing these questionnaires.

### *Information content*

The SFO collects data from the business economics (commercial) annual account, using paper questionnaires. The questions for large companies are much more detailed than those for enterprises with total assets of under 10 million guilders.

### *Timeliness*

In September of the year following the review year, provisional figures are published on the SFO with more than 10 million guilders; the publication level is more general than the definite figures. The definite figures for the entire SFO are published seven months later.

## **3. From VIS to SFO**

In spite of the differences between administrative concepts in the VIS and statistical concepts in the SFO, it appears that an SFO for enterprises with total assets of up to 25 million guilders based on VIS data is feasible.

The main reason for this is that in practice, for small enterprises the differences, which theoretically exist between population, units, information content and timeliness turn out to be smaller than expected. The existing differences are accepted as bias; for individual enterprises, translating from fiscal to commercial figures or consolidation of VIS units to figures at enterprise level is impossible without a glimpse behind the scenes. Moreover, the specification of the VIS annual accounts is absolutely inadequate to provide the desired information for enterprises with total assets of more than 25 million guilders;



with respect to enterprises with total assets of less than 25 million guilders, the only information lacking is the actual number of employees.

The calculation of SFO figures is based on the VIS, which contains the SFO target population. First the non-relevant VIS units are eliminated, such as units corresponding with enterprises with total assets of more than 25 million guilders, financial institutions and non-relevant legal structures. Estimates for the total population of declarations sent are then based on the average results of the declarations received.

#### **4. Technical aspects**

In order to use the data from an external administration like VIS, it is necessary to identify the units in that administration and to link these to the statistical units in the business register of Statistics Netherlands. This identification is not so simple as there is no uniform registration system for enterprises in the Netherlands. Each organisation, for example the Inland Revenue Service (IRS), the Chambers of Commerce and the social security boards, has its own business register with each its own coding system and maintenance procedures.

The central register of the IRS (*Beheer Van Relaties* or *BVR*) contains information on all taxpayers, natural persons (income tax) and corporate persons (corporate tax). All taxpayers are registered with a set of unique numbers to identify them in the different administrations, like VIS for corporate tax. These sets of numbers are the keys for the use of the fiscal data to compile business statistics.

In 1995 Statistics Netherlands started a pilot study to develop an identification procedure to trace the business units of the central IRS register, corporate persons and natural persons running a business, in its business register. Statistics Netherlands uses statistical units to compile its statistics. The statistical unit for the SFO is the enterprise, as a single legal unit or as a cluster of legal units. The IRS uses legal units or fiscal units, i.e. clusters of legal units which report as a whole for each cluster. The fiscal units may vary per tax (a unit for the corporate tax may differ from the unit for the value added tax). The starting point of the identification is the common unit: the legal unit. An extensive test on the units of the region Dordrecht showed that automated identification of 95% of the relevant units was feasible with 1% misidentification. It also became clear that identification took quite some processing time (5,000 units per hour of processing time), and that it would be possible to fit this in the regular production processes of the business register of Statistics Netherlands. The identification of the legal units is based on a comparison of legal name, trade name, legal address, mail address and local address and if necessary historical information. The identification is done by modular approach: first legal names and legal addresses are compared. If this does not produce any hits,

other information is used. The number of the module on which identification is based is recorded. In this way an indication of the quality of the identification is registered. The comparison of names and addresses was done by the method of trigrammes and LCS (largest common subsequence) score, and also based on correspondence in postal code and house number.

In December 1996 the initial identification of all business units of the central IRS-register in the Statistics Netherlands business register was realised. The percentage of successful identifications depends on the *fiscal activity status* of the units in the tax register. The results are shown in table 1.

**Table 1**  
**Link between legal units in IRS registers and Statistics Netherlands**

Fiscal activity status	Number of legal units in tax register	% of identified units
Corporate taxpayer	287,245	96
Subsidiary of corporate taxpayer	122,91	95
Corporate tax exemption	203,997	56
Not subject to corporate tax (incl. natural persons)	743,619	67
Overall	1,357,717	74

## 5. Thematic approach; critical success factors

### *Organisation*

With an innovative process such as making a secondary source applicable for the output of statistics, it is vitally important to use a thematic approach to effect this. The whole process starts with a thorough preparation, resulting in a detailed plan of approach. An implementation phase must also be planned to achieve a successful result, and the project organisation should not lead to any misunderstandings: one client and clear agreements as to the staffing of the project group. In aid of manageability, it is recommended that distinct phases be introduced with respect to eventual project results, and that the work should be carried out on an incremental basis.

### *Cooperation*

Close and open collaboration with the database owner, in this case the Ministry of Finance, is also a success factor. The collective objectives also contributed to this process.

### *Investment in research*

The importance of the willingness to invest in research should never be underestimated. From the beginning, an investigation in September 1994, to the implementation in April 1997, the *VIS to SFO* project has cost some seven to eight man-years. It should be noted that this investment goes further than the SFO alone; the knowledge gained with respect to the VIS can also be used to research the feasibility of related projects at Statistics Netherlands.

## **6. Consequences for the SFO**

### *Quality*

For statistical year 1992, in addition to the regular SFO, shadow statistics were compiled for companies with assets under 25 million guilders, based on the VIS. The VIS-based outcomes were substantially higher than those of the SFO; the structures of balance sheet and profit and loss account (individual items in percentages of total assets and turnover respectively) are comparable. The VIS level, however, links up better with the national accounts calculations than the regular SFO level. Due to the enormous increase in the number of declarations on which estimates are based, the precision of the VIS outcomes is far greater than the SFO outcomes. Table 2 reflects the results with their relative margins for two key variables.

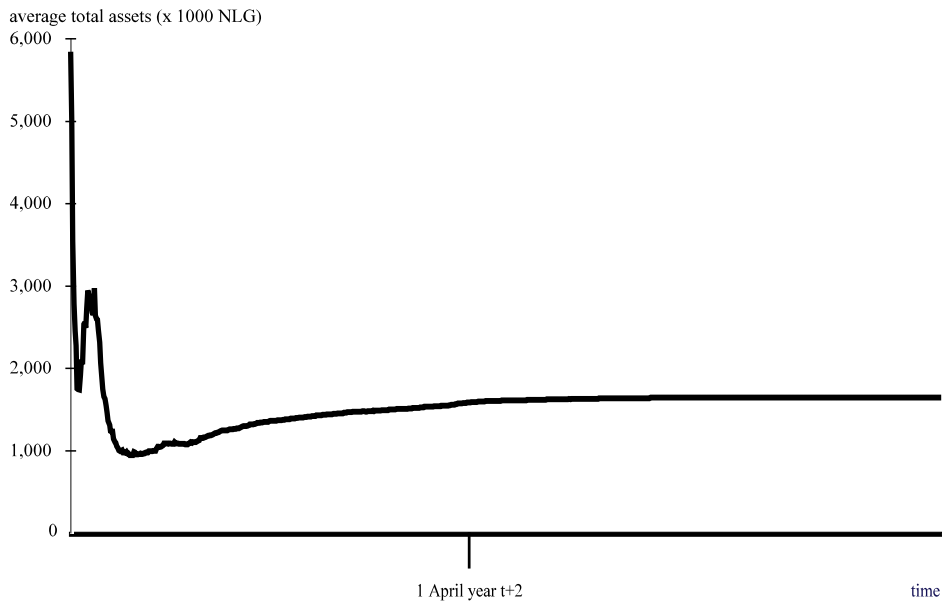
**Table 2**  
**SFO 1992 companies with assets under 25 million guilders; VIS versus traditional observation**

Variable	VIS estimate billion NLG	Relative margin	SFO estimate billion NLG	Relative margin
Total assets	229.2	0.006	186.9	0.20
Turnover	360.7	0.009	315.8	0.21

### *Timeliness*

The feasibility of the regular SFO planning using VIS data depends not only on the number of received declarations in the VIS, but perhaps more on the representativity or randomness of the received declarations. This is illustrated in Figure 1 by the development of the average total assets in the declarations received at a given moment, for the SFO 1992 of companies with assets under 25 million guilders. Variables such as turnover and net profit show a similar pattern.

**Figure 1**  
Development of the average total assets in the declarations received in time; review year  $t=1992$



The figure shows that the regular closing date for review year  $t$ , i.e. 1 April year  $t+2$ , remains acceptable. Although the estimated total assets will be too low, the figure is still reliable in the sense that it is calculated at a time when the course in time is stable. In the future, this underestimation will be tackled using several VIS years.

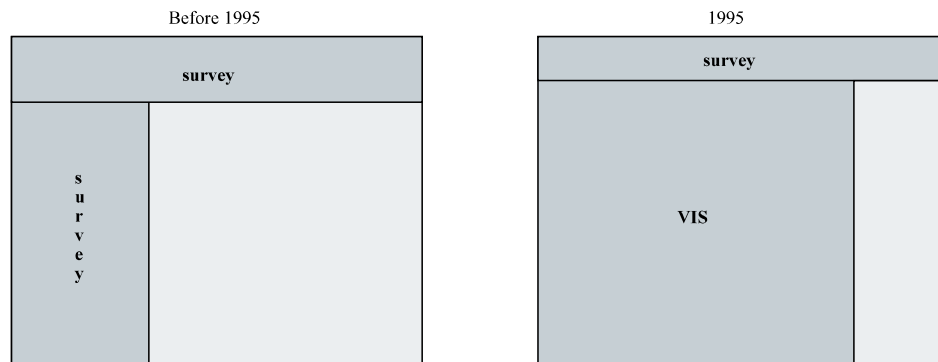
#### *Response burden*

With respect to the SFO response burden, using VIS means that some 7,000 enterprises no longer need to complete a questionnaire, which in turn means a saving of almost 10,000 hours of work, a reduction of over forty percent.

## **7. Consequences for the internal organisation**

VIS certainly makes a difference. Figure 2 illustrates the core of the changes. Fewer surveys were carried out in 1995: no more than 2,400 enterprises with total assets of more than 25 million guilders, while companies with assets of under 25 million guilders are no longer surveyed at all; their information is obtained from the annual accounts stored in the VIS database.

**Figure 2**  
**Change in SFO situation through the use of VIS in 1995**



This redesign of the SFO statistics has two important consequences for the internal organisation and the production process:

- Surveying, the traditional way of making statistics, has become less important. Although the impact of the survey on the results is still very high (75% of total assets), less time and manpower are spent on these activities, which dominated the production process before 1995.
- Adaptation to a completely different situation was necessary. New tools and new skills were needed: handling a huge database with the appropriate hardware and software; editing tens of thousands of records simultaneously, instead of editing individual forms; communicating with staff of another large organisation (Ministry of Finance); being much more dependent, instead of having personal control over the entire input and progress.

In the case of the VIS-SFO project, making statistics based on an existing central registration has ultimately resulted in a *regular fully computerised procedure*. Production of the SFO for enterprises with total assets of under 25 million guilders takes about a week from the moment the VIS database becomes available to the moment the statistical tables are completed. In real production time, however, it takes only around 10 hours. It is clear that this redesign has substantially changed the world of those working on the statistics. Now four people instead of seven run the SFO production process. Part of this work entails higher, and in many cases, varied skills. Better statistics are now being produced at lower costs and with a smaller response burden. On the other hand, the SFO now has to rely more on other departments, both inside and outside Statistics Netherlands.

## **8. Evaluation**

Secondary EDI offers real opportunities to reduce the survey response burden for the business community. The greatest gain will be in the case of small and medium-sized enterprises, where the response burden, in terms of number of forms per employee, is greatest (Van de Stadt 1995).

Following the SFO, a study has been started to examine whether the VIS can be used for other financial statistics, and whether the business register of Statistics Netherlands can be extended with fiscal information.

## **Reference**

Stadt, H. van de, 1995. Replacing business surveys by tax registers. Netherlands Official Statistics – winter 1995 (vol. 10), pages 28–32.

# Electronic commerce and EDI: a European perspective\*

Philippe Lebaube\*\* and Maarten Boon

*Statisticians may see the introduction of Electronic Commerce & EDI either as a threat or as an opportunity. A threat if the harmonisation of data resulting from the implementation of EDI in the information systems of economic operators means that information to be collected for statistical purposes is no longer available in the requested format or (for reasons of economy) cannot be collected by traditional means. An opportunity if Electronic Commerce & EDI could be used by statisticians to collect more by asking less. This could be realised by embedding statistical data collection mechanisms in Electronic Commerce services, EDI data flows or electronic form filling systems.*

*The present article describes the European initiatives and projects conducted by the statistics sector, particularly highlighting the roles of EDI standardisation activities and EDI projects. Conclusions are drawn from the experiences until now.*

## 1. Introduction

The information market is a basic component of our modern economy. The relentlessly increasing demand for information is no respecter of national frontiers, and the quality of the information required by users can be measured in terms of freshness, precision and representativeness. The various economic operators<sup>1)</sup> and government or semi-public agencies rely increasingly on public or private statistical systems to provide them with the key indicators they require for decision-making or the evaluation of options. The various statistical systems must meet these needs by collecting more information, making sure they can respond more quickly and, above all, adapt to changes in demand.

---

\* This article is a reworked version (at the request and with permission of the original author) of a paper presented by Philippe Lebaube at the G7 Electronic Commerce Conference.

\*\* Eurostat.

Recent developments in the field of informatics, and especially on the telecommunications front, have led to the introduction of Electronic Data Interchange (EDI) permitting the intercommunication of the information systems of the various economic operators in the course of the main phases of their commercial and administrative activities (production, distribution, invoicing, declaration, trading). Electronic Commerce and EDI are rapidly developing techniques in dynamic sectors of the economy like commerce, transport, car manufacturing, banking, chemicals, insurance, informatics. Government departments are also increasingly interested in EDI, either for their relations with enterprises (customs, taxation, social security) or for exchange of information with other administrative departments.

A structural change is taking place in the field of EDI applications, in which the mono-sectoral approach is giving way to multi-sectoral systems. The first applications of EDI were in fact confined to a homogeneous population of partners in very specific areas of activity. This can be illustrated by the 'old' lists of projects and the composition of the user groups concerned (car manufacturing, civil aviation and the chemical industry).

The current forward thrust of EDI is designed to secure the involvement of other partners from other sectors. Economic operators are keen to use the same EDI technique to communicate not only with their sub-contractors and suppliers but also with their bankers, transport services or government departments. The success of this multi-sectoral revolution depends on the extent to which the technical specifications for the EDI applications and Electronic Commerce services are based on common standards.

The number of software packages available on the market is increasing rapidly, and the main computer manufactures are already supplying EDI workstations and/or gateways. Due to the emerging telecommunications networks and the Internet phenomenon, Value Added Networks (VANs) are offering integrated services based on Electronic Commerce and EDI techniques. These new technologies will profoundly modify the structure of the information systems of the economic operators and their modes of intercommunication.

The progressive introduction of Electronic Commerce, EDI and the advent of the paperless society will make it increasingly difficult for statisticians to carry out their various tasks. This is because the economic operators will become more and more sensitive to the burden imposed by the collection of statistical data, since:

- the trend in the world of trade and commerce is decidedly in the direction of removal of restraints and elimination of administrative, legal and national barriers, and the



- sources of statistical observation in these fields can be expected to play a less positive role;
- the various economic operators, for their part, will tend to become increasingly reluctant to complete specific statistical questionnaires in an age when all their other productive, operational or administrative activities are highly automated.

In due course statisticians will have to reorganise their information collection processes with a view to their integration with the systems that govern the activities of the operators in the public or private sectors of the economy, in particular the business information systems of economic operators.

## **2. Electronic commerce, EDI and statisticians**

Economic operators' business information systems rely increasingly on EDI strategies as far as the streamlining of the operational relations with the different partners is concerned. The systems of direct statistical data collection must be based on the technical solutions adopted by the operators in their own operational environment.

*Therefore, from the outset, statisticians have to participate in the process of introduction of the new technologies so that statistical requirements are embedded where necessary.*

Ideally, statistical requirements should be taken into account in the design stage of Electronic Commerce services, EDI dataflows and the related EDI messages. These requirements relate not to the content of the information but also to how it is structured and coded. The information to be collected for statistical purposes would then be present in the business information systems of the economic operators.

The more the user groups advance in their message developments, the harder it will be to take statistical requirements into account. Investments in the design of messages that have already been standardised cannot be simply written off, and requests for modification at that stage are unlikely to be favourably received.

*Fortunately statisticians can make a useful contribution to the design and modelling of EDI applications because of their huge experience in the fields of methodology, data modelling and processing.*

Statistics is a structuring domain, by virtue of factors such as the methodological approach, the definition of variables and the use of classification systems and units of measurement. The know-how and experience in design and methodology of statisticians in the field of Electronic Data Processing (EDP) techniques could be drawn upon for the design of better EDI systems. Moreover, a considerable number of national and

international coordination and intra and inter-sectoral harmonisation projects have already been completed or are currently in progress in the field of statistical systems. The adaptation of business information systems to EDI techniques will inevitably impose a heavy workload on the specialists in modelling the information to be exchanged (and hence the information to be stored) in the business information systems. There will be a growing need for common nomenclatures and classification systems. On the basis of proven expertise and to their own advantage, statisticians may contribute to this market demand. Actions that spring immediately to mind are the promotion of techniques for data modelling and coding used for statistical purposes and, in particular, the utilisation of statistical nomenclatures in EDI messages. The latter is also a matter of vital importance for statisticians.

*And as a last – but not least – challenge: statisticians can benefit from the EDI technology to improve the quality, availability and user-friendliness of the statistical services they offer their users*

As far as the statistical systems themselves are concerned, the introduction of EDI and the interconnection of their information systems will bring the need for harmonisation and coordination into increasingly sharp relief. Projects are already in progress to promote the use of EDI technologies between national statistical administrations.

Similarly the use of Electronic Commerce services is under consideration to develop dissemination services of statistical products. The more, and the more accurate, statistical information the economic operators may receive in return in an appropriate way, the more they will be willing to supply the basic statistical information.

### **3. Standardisation and statisticians**

The widespread adoption of EDI systems has highlighted the need for standardisation of interchanges at an international level. The EDI standardisation process comprises:

- defining the syntax of the messages to be used for structuring and representing the information to be exchanged;
- listing and documenting all the messages that have been harmonised to meet the users' requirements;
- providing the information coding elements required by users;
- ensuring the follow-up of the pure standardisation process in relation to the EDI standardisation bodies.

From the outset, the tasks of standardisation in the EDI field have been entrusted to working groups specialised in the definition of measures to facilitate international trade<sup>2)</sup>.

The standard that is now taking the lead in the EDI field is the UN/EDIFACT standard<sup>3)</sup> which has been adopted at an international level. Although the North American countries use a different standard, ANSI-X.12, they played a key role in the definition of UN/EDIFACT and are now committed to convergence with this international standard. More than 250 UN/EDIFACT messages (e.g. invoices, customs declarations, orders for money transfers) are currently in the course of standardisation and evaluation. Every new year sees a number of new projects for using the UN/EDIFACT standard in other sectors, for example construction, the social sector, health, administration, chartered accountancy.

The coordination and consultation structures for these standardisation projects are already in place; substantial resources have been mobilised for the purpose and meetings of experts are organised by regional and international working groups. Work on the standardisation of UN/EDIFACT messages is being coordinated worldwide, the UN/EDIFACT directories are widely in use. Official declarations to public administrations using EDIFACT messages are now prescribed or recommended by some national administrations.

As far as statistics are concerned, the European developments are conducted under the aegis of the EBES (European Board of EDI Standards) Expert Group 6 - Statistics. This group is composed of the following working groups:

- WG1 – Exchange aggregated statistics,
- WG3 – Exchange of classifications,
- WG4 – Questionnaire data and raw data collection,
- WG5 – External trade statistics,
- WG6 – Balance of Payment,
- WG7 – Transport statistics.

A regular cooperation of EEG6 takes place with other EBES experts group, for example those working on customs, finance, transport, administration, commerce. International contacts have been established with other bodies involved in statistics such as the US Bureau of Census, US Bureau of Labour Statistics, Statistics Canada, Australian Bureau of Statistics, Japan Statistical office, IMF, World Bank, United Nation Statistical Office, OECD.

EEG6 has developed, maintained and documented the following messages:

*Generic messages*

- GESMES – Generic Statistical Message

A flexible message for the transmission of any type of multi-dimensional data or chronological series. GESMES supports the exchange of statistical figures together with

the associated meta-data (e.g. codes, footnotes, methodological notes, dataset structures). GESMES can also be used to request data. Many subsets of GESMES have been designed for similar types of exchanges: ECOSER, PRODCOM, BOPSTA, STATEM.

– CLASET – Classification exchange

Supports data exchange requirements for classification structures and definition, correspondence tables, codelists. CLASET also allows the exchange of the lifecycle information (updates and revisions).

– RDRMES – Raw Data Reporting message

Designed for returning the responses to questionnaires. The message, developed jointly with North-American and Australian.RDRMES, has been used successfully in the following declarative domains: VAT declarations, employment, UK steel industry, trading levels, excise details, material usage. RDRMES could also be used as a response message for acknowledgements and indications of errors.

*Declarative messages*

– External trade subset messages

The INTRASTAT regulation specifies that the electronic declaration must conform to the CUSDEC/INSTAT message, which is a subset of the customs declaration (CUSDEC). For declarations covering trade with non-Union countries the CUSDEC/EXSTAT message has been designed. The CUSDEC/INRES message was developed for replying and acknowledgement.

– Balance of payment reporting

The information needed for this reporting has been embedded into existing financial EDIFACT messages such as payment orders (PAYORD) and credit adviser (CREADV). A set of five messages has been designed to report statistical information and aggregated figures:

- BOPCUS: reporting of foreign payment by a bank on behalf of its customer;
- BOPBNK: reporting of a bank's own transactions;
- BOPDIR direct reporting by economic operators;
- BOPINF: information transmitted by a company to its banks when its account is credited;
- GESMES/BOPSTA: aggregated information exchanged between central banks, national statistical institutes and international organisations.

– Transport statistics reporting

Raw data are extracted from the IFCSUM message which is used within the maritime sector for reporting vessel calls at ports. The CUSCAR message is now under serious consideration by several transport operators for reporting vessel calls simultaneously to

port, customs and port health authorities as well as statistical administrations. The GESMES message is used for transmitting aggregated statistics between port authorities, CNAs and Eurostat.

EEG6 has developed data models to support the functions required by the EDI messages, and thus the implementation of some of these models in different exchange syntaxes (UN/EDIFACT, SGML, HTML). This shows the strength of the models because they are syntax independent and could be applied in the so called *open EDI environment*. It is probable that EEG6 will develop SGML and HTML versions of all generic statistical messages (GESMES, RDRMES, CLASET).

Efforts will be made to get the EDI messages used by the statistical sector well integrated into commercial software products and services by working closely with software houses and On-line Service Providers.

#### **4. Overview of European statistical initiatives and projects**

The European Statistical System (ESS) covers all statistics collection, processing and dissemination activities. These activities are described in the Statistical Programme <sup>4)</sup> of the Commission as adopted by the Council.

A set of European initiatives has been launched to promote the use of new technologies and telecommunications networks in the statistical domain. These initiatives are supported by Eurostat, the Statistical Office of the European Communities. They benefit from the development of Trans-European Networks, the IDA <sup>5)</sup> programme and the 4th Framework Programme.

##### *4.1 DSIS – Distributed Statistical Information Services*

DSIS <sup>6)</sup> is a framework for a wide range of technological activities supporting the functioning of the European Statistical System. These activities aim to ensure that technology is used effectively to benefit the community. Many partners are involved in DSIS activities: Eurostat, European commission directorates, European institutions or agencies, national statistical institutes of EU member countries, EFTA countries, and other administration providers, intermediaries or users of statistical information (e.g. customs, central banks, agriculture, information brokers, universities, professional unions, chambers of commerce).

The following priority statistical domains have been identified:

- \* national accounts
- \* balance of payments
- \* ECU/EURO statistics
- \* employment
- \* external trade
- \* industrial indicators
- \* PRODCOM
- \* price indices
- \* transport
- \* insurance
- \* agricultural indices
- \* health monitoring
- \* environment

The work programme, set out by the Statistical Programme Committee (SPC), laid down the following priorities.

#### 4.1.1 European reference environment

The European reference environment will provide a source of harmonised, validated and documented reference data from which disseminators can produce their statistical tables. Users of statistics will benefit from the access to multi-lingual data for which the methods, statistical concepts, classifications and dataset description and identifiers conform to agreed standards. The GESMES and CLASET messages have been recommended to exchange data and meta-data between reference environments. Implementation is underway at Eurostat and several European national statistical institutes.

#### 4.1.2 Raw data collection

This priority area encompasses the development of new collection methods of statistical data from economic operators using advanced technologies. The aim is to reduce the burden on the competent national administrations (statistical institutes) and the economic operators from which basic data are collected (about 15 million). The latter will benefit from simplified reporting mechanisms, embedded statistical requirements in business information systems or operational EDI messages and electronic form-filling systems. The benefits for CNAs will include improvements in the accuracy, completeness and timeliness of the raw data reported. Electronic transmission of data eliminates the data entry burden and increases the quality of statistics. The prioritised target group are the small and medium-sized enterprises (SMEs) that may not have access to large and flexible business information systems.

The trials of the use of EDI messages for reporting data in some statistical domains have demonstrated the viability of this approach. For the *balance of payments*, for instance, more than ten central banks already report to Eurostat monthly using the GESMES/BOPSTA message. Moreover, operational services have been open to companies in some countries to offer 'direct reporting'. The data interchanges may be made secure if requested. Trials are also being carried out in the field of *maritime transport statistics*.

Some other important raw data collection projects (EDICOM and SERT) are described in following paragraphs. The TELER project is outlined in another article in this issue.

#### 4.1.3 Multimedia and information highways

The set-up of dissemination environments for statistical information using multimedia technologies and information highways.

##### *Multimedia prototypes*

Eurostat and several national statistical offices are already applying the possibilities offered by the multimedia revolution and information highways. Photographs, charts, sound, video sequences and animation can be used to illustrate statistical data, making them more attractive, understandable and user-friendly. A whole range of products are already produced regularly: multimedia catalogues, CD-ROM, on-line multimedia pages on Internet and information service providers, information kiosks, etc. The Eurostat multimedia kiosk has already been on display at several exhibitions.

##### *The DSIS information resource centre*

Based on Internet technology, the DSIS information resource centre or DSIS-IRC will offer the following services for each of the hosted *interest groups*:

- \* document and software libraries;
- \* on-line conferencing (*cyber meeting*);
- \* newsgroup facility;
- \* directories of operational information such as project and meetings information, who's who, frequently asked questions;
- \* on-line services such as data collection services, reference environments, EDI show room, benchmarks;
- \* shop services where customers can order statistical products;
- \* global search facilities.

For each *interest group* user access and security mechanisms will be arranged according to requirements. Shop services will experiment with electronic commerce results (e.g. secure transaction).

The DSIS-IRC is currently being piloted at some national statistical administrations. Once the resource centre is mature and reliable it will be integrated into the www architecture of the European Commission and cross-linked with the www services of national administrations. The creation of an EEG6-site is foreseen for 1997, as part of the DSIS-IRC for EEG6 and its subgroups, to make documents, results, benchmarks, EDI showroom and products available to a wider audience.

#### *Support to the datashop network*

The support to the datashop network will be part of the trials of the shop services from the DSIS information resource centre. The datashops are established as service companies, whose mission is to provide fast and efficient services to the customer. Each shop is a joint venture between a local service company and Eurostat. The shops are located in Brussels, Luxembourg, Rome and New York; new shops are planned in Germany and the UK.

#### *Some relevant www sites*

IDA: <http://www.ispo.cec.be/ida/ida.html>

Eurostat: <http://www.cec.lu/en/comm/eurostat/eurostat.html>

NSI's: <http://www.cec.lu/en/comm/eurostat/guide/natstat.html>

EBES: <http://www.ebes.cenclbel.be>

#### 4.1.4 Design and integration of EDI messages

As described in section 3, the design and integration of statistical EDI messages has made considerable progress.

#### 4.1.5 Integration with trans-European networks

All partners involved in the European Statistical System should have access to 'good commonly available telematic services'<sup>7)</sup>. This availability is on the critical path of most actions of the DSIS work programme. The so-called *DSIS logical network* should provide the backbone between the various partners using the trans-European networks. Easily accessible and affordable telecommunications will improve the inter-operability of information systems and encourage cooperation at all levels. The work programme therefore aims to identify, promote and integrate tools, products and services which provide a sound technological base to develop distributed statistical information services on top of existing trans-European networks and telecommunications infrastructures.

The carrier might be one of a number of technical means, such as X.400, Internet, file transfer protocols, client/server architectures. The use of TESTA<sup>8)</sup> services is being considered. A draft 'Memorandum of Understanding for Access Service Providers' specifying the various types of services expected has been elaborated and validated with major Internet service providers (ISP), access service providers (ASP) and on-line service providers (OSP). The basic requirement is to provide all or some European statistical actors with technological resources (e.g. services, tools, etc.) for setting-up, developing and operating on-line services to support the collection and dissemination of statistical data processes.



## 4.2 EDICOM<sup>9)</sup> and EXTRACOM<sup>10)</sup>

With the abolition of customs declarations for intra community-trade in 1993, the foreign trade statistics were deprived of a traditional source of information. The only alternative was to get the statistics directly from the trading enterprises. The INTRASTAT system novel in that it collected the import and export data directly from enterprises which have to provide the relevant national authority with a monthly declaration. All EU member states have made a considerable effort to realise the implementation of INTRASTAT. Supplemented by Community measures, these actions received support under the EDICOM project. Moreover a comparable project was started for the extra community trade, under the name of EXTRACOM.

The SLIM (Simpler Legislation for Internal Market) initiative is looking at ways to simplify legislation in four areas of the internal market in order to reduce burdens on business. One of these areas is the INTRASTAT regulation. Actions investigated are: extended use of EDI and telematic networks, simplification of reporting classifications and of the INTRASTAT regulation.

The following achievements have already been made under the EDICOM/EXTRACOM umbrella.

The *electronic questionnaire* incorporated in the IDEP<sup>11)/CN8<sup>12)</sup></sup> software, developed under responsibility of Eurostat, allows enterprises to compile their monthly declarations. The CN8 component of the software contains the entire product classification and allows searching for the appropriate coding of products using different routes. Eurostat places IDEP/CN8 at the disposal of the competent national administrations which provide this software free of charge to the enterprises. The package is currently distributed in ten countries; and the number of users is probably in excess of 20,000. The Netherlands and Germany use the IRIS software developed by the Statistics Netherlands (over 10,000 users)<sup>13)</sup>. Several software houses across the EU are offering equivalent products. In several countries the commercial software needs to be certified by the respective competent national administrations before it may be used.

Several *EDIFACT messages* for INTRASTAT and EXTRASTAT have been developed and already implemented (see section 3). The security and integrity of electronic transmission using EDI and telecommunications networks have been tested with success.

A large proportion of INTRASTAT declarations are still sent to CNA's on diskettes, tapes or paper. *Disk/fax systems* have been developed to read these large quantities of diskettes and tapes and to convert the data. Fax messages can automatically be sent to enterprises whose declarations were not accepted due to formatting errors, for example. *Optical character recognition systems* are also used to process paper declarations.

The *telecommunications projects* aim to permit the transmission of EDI declarations directly to the CNA's by using telematic networks. Replies and acknowledgements of receipt may be returned to enterprises. The IDEP software includes such an interface.

Once collected and harmonised, the statistics on intra and extra-community trade are available at Eurostat in a single reference environment together with classifications, conversion factors and indices. The *COMEXT systems* are used mainly by Commission departments. Dissemination outside the Commission takes place via publications, but mainly via the *COMEXT CD-ROM*. Projects are underway to benefit from multimedia technology and telematic networks.

The *measures taken at national level* take a different form, since they are a feature of the organisations chosen by the member states on the basis of the subsidiarity principle.

#### 4.3 SERT – *Statistiques d'Entreprise et Réseaux Télématiques*<sup>14)</sup>

The main purpose of the SERT project is to automate the production of statistical data for business surveys, with a simpler and quicker response to the surveys, lower costs, especially for small and medium-sized businesses (SME), which are hit relatively harder by these constraints. By analysing the requirements of both the administrative data collecting agencies and the organisation of the information system of the enterprise (BISE<sup>15)</sup> and SISE<sup>16)</sup> concepts), the project aims to create a real and open market for software developers and software houses in the field of EDI packages for data collection:

- avoid the proliferation of closed EDI systems and the widespread incompatibility which this entails;
- promote the creation of trade EDI systems that meet the needs of the users, particularly the small and medium-sized enterprises (SME);
- increase awareness of the European telematic equipment and service industry to meet these user requirements;
- support the use of common EDI standards, such as EDIFACT, SGML;
- increase the awareness of the various administrations that are collecting data from enterprises in Europe.

More than 19 European EDI pilot projects have been surveyed and analysed. In 1996, several pilot projects were conducted in Europe to validate the SERT approach. One of these was EDIVAT, a project concerning the transfer of VAT data (Value Added Tax) from chartered accountants to tax authorities. Statistical data are derived and transmitted to the national statistical institutes which compile business related statistics. The RDRMES message was used successfully in pilot projects conducted in Belgium and the Netherlands. Other countries, like the UK are considering joining in this project.

#### 4.4 4th Framework Programme for Research and Technological Development

The statistical sector is involved in R&D work in order to evaluate the potential of new technologies for the various phases of statistical processes from data collection to data dissemination<sup>17)</sup>. One of the statistical projects is TELER<sup>18)</sup> (Telematics for Enterprise Reporting).

The EUROFER-UK project has demonstrated that administration costs can be reduced, financial control can be tightened, pricing strategy can be based on better commercial information and more accurate statistical data can be sent to the ISSB (steel union). The project also demonstrated that EDI/EDIFACT is perceived as costly and complex, enterprises want control over release of information and better feedback. It may be difficult to justify investment in EDI just to improve statistical requirements.

### 5. Conclusions

The technical lessons learned from the various initiatives and project pilots can be summed up in five 'golden rules' for the success of initiatives making use of electronic commerce, EDI and telematic networks:

1. commitment of the major players, the 'hub' concept, to encourage the user communities and reach the critical mass;
2. well documented requirements (for example via EDI messages) and coherent implementation of technologies across the various sectors;
3. availability of standardised technical solutions (e.g. EDI messages) designed to cover user requirements;
4. off-the-shelf business software applications with well-integrated, affordable and well-supported EDI technology;
5. interoperable, affordable, easily accessible, secure and auditable telecommunications networks.

The administrations have the initiative and a clear mandate on rules 1, 2 and 3. The administrations have to specify, document and prototype their requirements on rules 4 and 5. The IT industry must have the initiative or a leading role on rule 4. The network providers and network operators have the initiative or a leading role on rules 4 and 5.

## Notes

- 1) For the purpose of the present report, an *economic operator* is defined as a private or public enterprise, a socio-professional organisation, a trade association in a specific branch of economic activity, etc.
- 2) UN/ECE-WP4 – Working party on facilitation on international trade procedures.
- 3) ISO 9735 EDIFACT: Message syntax;  
ISO 7372: Trade data element directory.
- 4) Currently 1995–1997, the 1998-2002 Work Programme is in preparation.
- 5) IDA – Interchange of Data between Administrations – European Commission-DG3 programme.
- 6) DSIS is part of the IDA programme.
- 7) As defined in the IDA Architecture Guidelines.
- 8) TESTA – Trans-European Services for Telematics between Administrations, IDA project.
- 9) EDICOM: EDI for Commerce – Eurostat initiative.
- 10) EXTRACOM: The EXTRACOM initiative consists of a set of coordinated actions relating to foreign trade Statistics in connection with the use of telematics. EXTRACOM is part of the IDA programme.
- 11) IDEP: Intrastat Data Entry Package.
- 12) CN8: Combined Nomenclature 8 digits.
- 13) Described elsewhere in this issue.
- 14) SERT is part of the IDA programme.
- 15) BISE – Base d'Information Statistique de l'Entreprise.
- 16) SISE – Système d'Information Statistique de l'Entreprise.
- 17) For details see also www pages:  
<http://europa.eu.int/en/comm/eurostat/research/intro.html>.
- 18) Described elsewhere in this issue.

# EDI, the future

Wouter J. Keller

*Although the developments in EDI, described in this special issue, have only just begun, it is already possible to draw some conclusions. Furthermore it is possible to indicate where it all might lead to. It is clear that the developments in EDI will not only have consequences for the way we communicate with our respondents and our customers, but also for how we organise our statistical process.*

## 1. Introduction

In the year 2010, when everybody regularly uses the electronic highway to communicate, shop and work, it might be very difficult to get respondents to fill in paper questionnaires. For personal, household and simple establishment interviews, we will use CAPI (Computer Assisted Personal Interviewing), CATI (Computer Assisted Telephone Interviewing), and some form of CASI (Computer Assisted Self-Interviewing) on the Internet, where respondents fill in simple forms. For complex establishment surveys, however, we need an alternative to mail-out paper questionnaires, like EDI, which connects business systems electronically to statistical ones.

From the outside, it might look as if the problems involved are mainly technical. As we shall see, however, much more effort will be needed in the conceptual and organisational areas to make this worthwhile. In this article, I shall discuss these and other implications of EDI for future statistical processing.

## 2. Technological developments

The Internet (or World Wide Web) is the source of many revolutions in Information Technology (IT) these days, and will continue to be an important source of change in the future. Instead of concentrating our IT strategies on the internal processing within statistical institutes, as we used to, we shall be forced to open up to the external world. Just as manufacturers (e.g. of cameras) and service providers (e.g. travel agencies) will set up electronic links with their suppliers (for lenses or hotels) and customers (buying on the Web), we need to connect to the external world too. The success of this external

IT orientation of statistical institutes will, in my opinion, determine the future of our profession. As we put it at Statistics Netherlands: we shall have to become a junction on the electronic or information highway if we want to survive.

When booksellers like Amazon, computer suppliers like Dell, and software shops like ISN already ring up more than a million US dollars a day in sales over the Web, people will expect us too to be present on this electronic highway in the future. And since our main suppliers are information providers (our respondents!) and our main customers are information buyers, we can deal with information both on input and on output. In other words, our input and output materials are in essence *bits*, and therefore very easily handled on the Web, without any of the logistical problems involved in shipping books or PC's. At the same time, once all information is presented in the form of bits, we are able to tap more information without burdening the suppliers. As we shall see here, we might even go out looking for information on the electronic highway, irrespective of its location or origin, in order to describe the world around us in statistical terms, as we are supposed to do now.

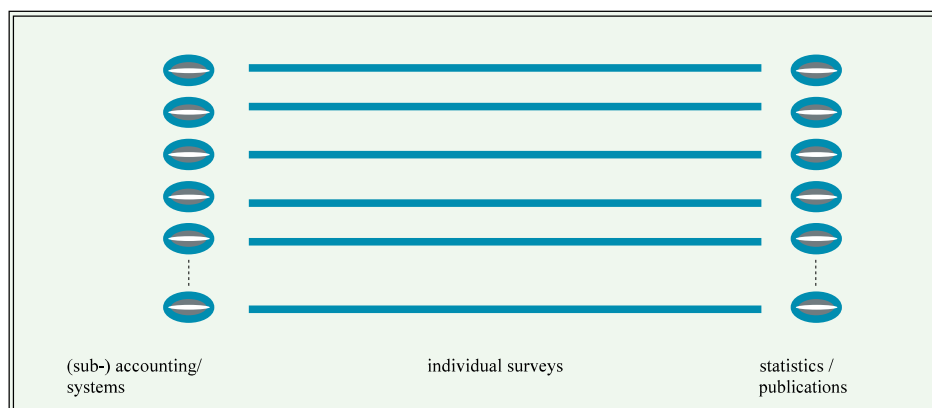
While in the past statistical data were found mainly at statistical institutes, today statistical information is increasingly used in business. Data-warehouses - collections of, for example, customer information, sometimes of immense proportions, provide modern management with the means to focus on interesting new markets and other strategic issues. Electronic facilities like Smartcards and loyalty programmes like the Dutch AirMiles will only add more data to these warehouses. When the large volume of consumer transactions really takes over the Net, most transactions will be registered on the Web, leaving many electronic traces in as many data-warehouses. All this machine-readable information is becoming available not only to the institutions involved, but potentially also to statistical services.

In order to handle individual statistical data on the Web, there is a need for encryption and identification (or authentication). While pollution by industry was the problem of the last decades, privacy and security on the Web is becoming today's electronic counterpart. However, with emerging technologies like Smartcards, SET (Secure Electronic Transactions) and institutions like TTP (Trusted Third Parties, e.g. banks), within a few years it will be possible to securely exchange confidential data and to guarantee the identity of respondents and statistical institutes involved. However, this will not imply that there will be no privacy problems in the future: with better access to more information about individuals and establishments, the need for privacy protection will become more and more important.

### 3. Focus on sources, not surveys

In the past, each department within the statistical institute handled its own surveys, questionnaires and publications. As a result, a medium-sized enterprise in the Netherlands received sometimes as many as 40 questionnaires and forms a year from us. This is typical for traditional companies, which are set up according to strict functional hierarchies and Taylor-like specialisation, each department with its own input, throughput and output stream. As a result, the traditional institutes reflect the so-called *stovepipe* organisation with little or no coordination or integration between different departments.

Figure 1



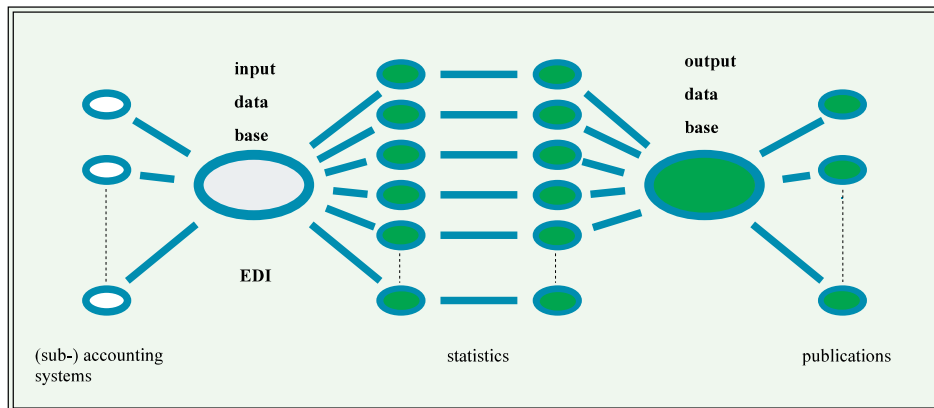
A typical example is the traditional supplier of insurance services, where each department sells its own policies for car, life, house insurance etc. with several steps between marketing and insurance, getting price quotes, handling paperwork and sending out policies to customers. These days, people want to phone or electronically contact their insurance suppliers and expect to be able to arrange any type of insurance interactively immediately in one session, without being bothered with all the traditional departments and steps. In other words, people do not want stovepipe organisations but direct, one-stop shopping facilities. This requires integration and coordination between stovepipes, and that is exactly what Business Process Redesign (BPR) is aiming at.

In statistics, we see this trend most clearly when we put our statistical information on the Web. In the Netherlands, we have initiated a statistical warehouse on the Web (called StatLine) where all our public information is made available. Based on the traditional stovepipe organisation, a great many different publications can be found in this database. A customer searching for information on unemployment, for example, is faced with more than a dozen different publications to choose from. Looking in these publications, the customer will see different definitions, different time and regional classifications,

different related variables, and so on. In other words, the electronic format, where comparisons are so much easier than with the traditional paper publications, demands coordinated data over publications, or even better, one comprehensive publication (database) where all topics are coordinated with respect to definitions, time, and regional classifications and so on.

On the input side of a statistical institute too, we see traditionally the same specialisation: each department conducts its own surveys and sends out its own questionnaires. It reminds us of the way the Dutch Internal Revenue Service used to collect taxes: each tax department (VAT, income, wealth, corporate, etc.) had its own processing and 'customers' were faced with many different requests from different departments. Now, the Dutch IRS is organised according to its *customers*, and one organisational unit handles all my tax returns. This *source* instead of *destination* oriented collection is the way to go for statistics, especially as the information for different surveys often originates from overlapping sources like financial, logistic and payroll administrations. When these sources comprise electronic (machine readable) information, the time is ripe for EDI: we should collect as much as possible electronically from one source and distribute the information to different survey departments later at our institute (see Figure 2). A similar reasoning holds for the output side, where we should integrate our statistical information into one database before distributing it to different media.

Figure 2





#### **4. Catch what you can**

We have seen that EDI allows us to concentrate on the respondent sources of surveys instead of the surveys themselves. We could go a step further and try to find information about respondents, independent of the surveys, focusing on sources anywhere. This line of reasoning will bring us to a strategy to collect relevant information from tax and social security registers, household and housing registers, yes even from distributor, wholesale and outlet databases, in order to build a picture of reality, by combining electronic sources at any time from any place. This is the vision behind Statistics Netherlands strategy (called *CBS 2000: junction on the electronic highway*) where electronic information from several sources will replace the traditional paper and pencil surveys.

This 'catch what you can' strategy also has implications for today's paper questionnaires. We should try to concentrate on what is available in the respondents' business accounts, instead of focusing on the concepts needed for, say, the national accounts. Asking for concepts which are not registered in respondents' administrations is not very useful and might lead to wild guesstimates, unrelated to other, hard, information. Therefore, it is of the utmost importance to collect *available* information and to reconstruct reality at the statistical institute using various sources.

In order to concentrate on sources instead of on surveys and to put the 'catch what you can' approach into practice, we need better methodologies to estimate statistical concepts like those for the national accounts. Record linkage, imputation, synthetic estimation and advanced post-stratification techniques are called for to combine the information from different sources. And in view of the huge electronic registers available today (e.g. tax and social security registers and wage administrations), we shall often stop sampling, as it will be easier to collect everything electronically than only a part on paper. Therefore, we have to develop new techniques to combine comprehensive (i.e. non-sampled) registers with sampled information, much like traditional post-stratification but with much richer comprehensive databases.

#### **5. Business processing redesign at statistical institutes**

The picture outlined above entails a radical reorientation of the statistical processes at statistical institutes. On the one hand, we have to connect to the outside world through integrated databases for electronic data collection and dissemination. As we have seen, this entails an orientation towards sources of information other than those used in the past. Individual surveys and publications (from the traditional 'stovepipes') are replaced by coordinated collection and dissemination through electronic databases at our input and output side. And not only in our external orientation, but internally, too, we shall see

dramatic changes. For example, we need new techniques for editing huge electronic databases, and when I say editing, I don't mean checking for adding-up errors or other incidental glitches. What we shall be looking at tomorrow are much more structural changes, for example when enterprises change their bookkeeping practices, their information systems, or their tax registration. We shall need intelligent systems which are able to find semantic instead of syntactical changes in particular, and we might need advanced techniques like artificial intelligence to do so, taking into account business rules per respondent.

But new methods are not only needed for the post data-collection processes. As with CAPI and CATI, the emphasis might change from editing after the survey to designing electronic questionnaires beforehand. If we focus on one source instead of one survey, we should try to define our electronic questionnaire (the EDI message) in such a way that we are able to provide the right information for many different surveys from that single source. Here, the conceptual differences between survey variables are at stake and effectively a huge effort of statistical coordination will be required to get all our various concepts across all the surveys in line. Again, the focus is much more on semantics instead of on syntax. As we have found out in the Netherlands, in this respect the specific character and structure of electronic messages helps more than the standards and paper rules of the past. The same holds for the output side, where customers on the Internet will demand consistent information from a statistical institute, independent of the survey or department it comes from.

As a result of these external changes, the internal processes at our institutes will have to change dramatically. The traditional stovepipes, dealing with paper surveys and paper publications of their own, will have to change to knowledge areas. For these areas, it is not the processing but the *knowledge* about certain aspects of society (labour, health, employment, etc.) that is the core business, together with new methods to make advanced estimates from different sources of information. They also define the input and output concepts and the necessary semantic translations between them for their knowledge areas. This should allow the input and output divisions to coordinate collection and dissemination as described above.

## **6. Meta data: syntax, semantics and knowledge**

The Internet was able to become reality because we were all using the same standards in electronics: protocols like TCP/IP, HTTP, and SMTP enable us to exchange information on a larger scale than has ever been possible. These standards refer to, among other things, the *syntax* of the messages we exchange. The next revolution will be one step up in abstraction: instead of syntax we shall need standards in *semantics* and, ultimately, standards in *knowledge*. An example is StatLine, our output database, comprising data

and meta-data. The first requirement for meta-data is syntax: we should know that the third field of a record starts at position 5, has length 4 and contains 4 characters, say '1997'. Then, the semantics should tell us that this should be interpreted as, 'year'. Lastly, there are some rules about 'year', based on the knowledge that 'year' refers to the concept of *time* and *time* has to obey certain rules (*knowledge*). One of these rules, for example, is that some elements are more recent than others, depending on the time and that we can see historical trends by watching the same variable over time. So, if someone formulates a query (the consumer price index in the Netherlands, say) it is reasonable to assume that they want at least the most recent figure and for that we need to know which dimension in the query reflects the concept 'time' and which direction of that dimension points to the future. This knowledge is more than simple syntax (format) or simple semantics (the meaning of 'year'), since it describes the nature of the dimension 'time' and the things we can do with it.

Future standardisation will involve more than syntax and more than (simple) semantics: it will concentrate on knowledge or *rules*. Just as tax authorities standardise the rules for depreciation in business, we shall need to cope with EDI messages which, in the future, will know about these different rules used by different institutions. So, in the future we shall build software for data collection and data dissemination that knows about the rules behind the concepts. This will require enormous effort, as each institution (respondent, data user) will have its own rules, alongside some standard rules imposed by suppliers, customers, governments or others (e.g. Internet).

## **7. No rules but tools**

There is a lesson to be learned from recent technological developments. Once a tool is successful, its success dictates its usage and everything needed for its usage. Internet is a successful data communication tool, and it now dictates the technical syntax we use for data communication. No official influence, no rule is needed to make us use the right syntax, it just does not make sense not to. The successful tool sets the rules. Statisticians can learn from this.

First they should think twice before opposing widely existing practices – above the definition of depreciation was mentioned – even if they are convinced that their own views are superior. If they want to change existing practices they should make certain that their position in 'the market' is strong enough. For this reason it will be advantageous to seek cooperation with another party with a common interest like the internal revenue service, who are also looking at ways to reduce the administrative burden.

The second lesson is that an objective is more easily achieved by means of an attractive tool than by means of a severe rule. This applies both inside and outside the statistical office. Clearly it is easier to use this approach inside the own institution. But outside, too, much can be achieved. If an attractive tool is offered for the collection of statistical data, one that makes the job of answering the statistical questions (whether obliged by law or not) easier and cheaper, there is a better chance of better quality and willingness to provide the required information, even if this takes some effort in the conceptual translation. This argument applies specifically to primary EDI. Providing an adequate and attractive tool that makes work easier and cheaper will often be more effective for getting the right results than issuing all kinds of detailed rules on what we expect people to do.

Of course, if statistics is desperately in need of very specific information, then maybe a tool cannot compensate for the extra effort needed to produce that information. A cost-benefit analysis should be made. We should ask ourselves how it is possible that statisticians describing reality need information which is apparently not of interest to the actors they are describing? The introduction of the EDI tools is a good opportunity to evaluate the necessity of our questions. One example is the simplifications now proposed on the intra-community trade statistics covered by the IRIS tool. Even so, sometimes rules will be needed to support the tool.

Furthermore, and this is also observed elsewhere, the success of a tool may turn into an impediment to change. We see this too in the EDI projects described here. A certain investment is always needed to define the relationship between the respondents' data and the data needed by Statistics Netherlands. Any change, technical or conceptual, will necessitate additional investment, and the extra benefits of this change should exceed those costs. EDI will probably lead to fewer changes in our questionnaires, a favourable prospect as far as the administrative burden is concerned.